U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

EFICIÊNCIA ENERGÉTICA E CONFORTO

Aplicação de técnicas de "*Data Mining*" à avaliação do ambiente interior

DANILO JOSÉ LEITE PINTO

Dissertação submetida para satisfação parcial dos requisitos do grau de

MESTRE EM ENGENHARIA CIVIL — ESPECIALIZAÇÃO EM CONSTRUÇÕES

Professor Doutor Nuno Manuel Monteiro Ramos

Professora Doutora Maria de Lurdes de Oliveira Simões

JUNHO DE 2016

MESTRADO INTEGRADO EM ENGENHARIA CIVIL 2015/2016

DEPARTAMENTO DE ENGENHARIA CIVIL

Tel. +351-22-508 1901

Fax +351-22-508 1446

✉ miec@fe.up.pt

Editado por

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Rua Dr. Roberto Frias

4200-465 PORTO

Portugal

Tel. +351-22-508 1400

Fax +351-22-508 1440

✉ feup@fe.up.pt

🌍 http://www.fe.up.pt

Aos Meus Pais

Ao Dário

À Juh

Quem encontra prazer na solidão, ou é fera selvagem ou é Deus

Aristóteles

AGRADECIMENTOS

Com a conclusão desta dissertação não posso deixar de agradecer a todos aqueles que direta ou indiretamente permitiram que este momento se torna-se realidade.

Agradeço aos meus orientadores, Professor Nuno Monteiro Ramos e Professora Maria Lurdes Simões, pelo lançamento do tema inovador aqui desenvolvido, pela disponibilidade e apoio incondicional ao longo deste trabalho.

Agradeço a todos os meus amigos que, direta ou indiretamente, me ajudaram a superar as diferentes dificuldades tornando este momento possível.

Agradeço aos meus Pais, Rui Pinto e Rosa Leite, ao meu irmão, Dário Pinto, e à Joana Lopes por nunca me deixarem incompleto, fazendo com que este sonho se tornasse realidade.

Por fim agradeço a todos aqueles, presentes e não presentes, por fazerem de mim uma pessoa mais completa todos os dias, impedindo que me afaste do caminho certo.

RESUMO

A Engenharia Civil é uma área em constante evolução e graças à sua grande abrangência permite que cada vez mais sejam explorados novos caminhos. *Data Mining* (DM) é um caminho de prospeção que apresenta grande potencial. Neste trabalho apresenta-se um cruzamento entre estas duas áreas de conhecimento, *Data Mining* e a Engenharia Civil.

O conceito de *Data Mining*, visa essencialmente à análise de conjuntos de dados fornecidos por experiencias variadas. É inovador na área da Engenharia Civil e, neste caso, foi aplicado à informação recolhida na análise de um conjunto de casas finlandesas, em que se estudou a relação entre humidade relativa e temperatura interiores, isolamento e o número de renovações horárias na influência do consumo energético.

Recorrendo a uma base de dados inicial aplicam-se as várias ferramentas disponíveis na realização deste processo. É realizado um tratamento inicial da informação e, devido ao grande potencial da análise de componentes principais, é conseguida uma redução significativa da dimensão da amostra em estudo. Os padrões encontrados, objetivo principal do processo de DM, são o resultado dos métodos de análise de *clusters* e do estudo da árvore de decisão construída. O método da análise de *clusters* é dotado de uma grande capacidade de aglomeração dos diferentes dados em conjuntos que sejam padronizados. A árvore de decisão consiste num método classificativo extremamente útil na identificação dos padrões existentes, permitindo uma fácil interpretação.

Este trabalho permitiu demonstrar a potencialidade de *Data Mining* e a sua aplicabilidade na Engenharia Civil. Foi possível verificar a influência do isolamento e RPH no consumo energético de diferentes habitações.

Concluiu-se que *Data Mining* apresenta bons resultados como um processo capaz de analisar grandes bases de dados e apresenta potencial na análise do conforto interior. DM fornece resultados que vão certamente ajudar o setor a responder aos desafios do futuro, tornando-o mais eficaz, ecológico e dando-lhe só ferramentas mas também capacidade de moldar um mundo melhor.

PALAVRAS-CHAVE: *Data Mining*, Conforto Interior, Eficiência Energética, Análise de *Clusters*, Árvore de Decisão.

ABSTRACT

Civil Engineering is an area in constant evolution and, thankful to its wide scope, allows news paths to be explored in name of the evolution. Data Mining (DM) is one of these paths of prospection that presents great potential. In this work is presented the intersection between this two areas of knowledge, Civil Engineering and Data Mining.

This paper presents itself, in the eyes of the Civil Engineering, the innovator concept of Data Mining. The process and the tools that belong to the different phases are here explored. This concept is applied to a group of Finnish detached houses, studying the relationship between the relative humidity of the interior and internal temperature, insolation and the ventilation rate, with the energy efficiency.

To the initial database were applied several tools available in this process. In the first phase is applied the initial treatment of the data to further application of dimension reduction methods. Principal Components Analysis revealed itself a powerful tool in dimension reduction of the data. The patterns found, which are the DM main objectives, were result of the clusters analysis and the construction of a decision tree. The first one is powered with a great capacity of agglomeration of the different data in different groups. The decision tree it's a classification method extremely useful to identify the data patterns, allowing an easy interpretation.

In this work was demonstrated the capabilities of Data Mining and it's applicability to the Civil Engineering. Was possible to see the influence of insolation and ventilation rate in energy consumption of different houses.

This study concluded that Data Mining presents great results in the process of analyzing big databases and shows the potential to analyses and make conclusions about internal comfort. DM also realese valuable information that can be applied in future constructions, improving even more the sector of Civil Engineering.

KEYWORDS: Data Mining, internal comfort, energy efficiency, cluster analysis, decision tree.

ÍNDICE GERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

SÍMBOLOS E ABREVIATURAS

ACP - Análise de Componentes Principais

DM - *Data Mining*

GUI - Interface Gráfica do Utilizador

CLI - Interface tipo Linha de Comandos

KDD - *Knowledge Discovery from Data*

SVM - Support Vector Machine

C - *Cluster* Modelo

r_{ij} - Correlação

s_{ij} - Covariância

$|p - p'|$ - Distância entre os Objetos p e p'

D - Distância Euclidiana

d_{max} - Distância Máxima

d_{avg}- Distância Média

HR - Humidade Relativa (%)

I_1 - Isolamento do Tipo 1

I - Matriz Identidade

S - Matriz Modelo

$\overline{x}$ - Média Aritmética

$P(B|A)$ - Probabilidade Condicionada de A na Ocorrência de B

$P(A \cup B)$ - Probabilidade da Reunião do Acontecimento A com B

V_1 - RPH do Tipo 1

RPH - Taxa Nominal Horária de Renovação de Ar Interior (1/h)

T - Temperatura (Cº)

u_i - Vetor Coluna

λ_i - Valor Próprio da Matriz S

1 INTRODUÇÃO

1.1. ENQUADRAMENTO GERAL

O crescimento tecnológico que se tem vindo a verificar ao longo dos anos incutiu na sociedade uma grande febre de recolha e armazenamento de dados. Os dados são armazenados com os mais variados objetivos, sendo que o principal é o seu estudo estatístico. Este tipo de estudos permitem transformar diferentes acontecimentos em números e gráficos que permitem uma melhor interpretação dos mesmos. Este tipo de análises podem ser consideradas como análises de primeira linha, isto é, o objetivo primário da recolha e armazenamento de informação. Após essa utilização pode iniciar-se o papel de *Data Mining* (DM), às enormes quantidades de dados agora inutilizados, que entra como análise de segunda linha. Este processo tenta encontrar padrões específicos nos diferentes conjuntos de dados. Este facto traduz-se como uma nova fonte de conhecimento capaz de descrever relações que existem mas que por vários motivos estão ocultas. Estas relações podem ser suficientemente fortes para que sejam muito úteis na tomada de decisões [1,2].

Já na perspetiva da Engenharia Civil esta tem-se mantido à margem do armazenamento de informação na maioria das suas áreas. Contudo este setor como todos os outros é gerido pelo grau de evolução de que a humanidade está sujeita. E por isso cada vez mais se tentam encontrar soluções capazes de rentabilizar ao máximo todos os processos construtivos, projetos ou manutenção de edifícios, bem como a sua utilização.

Este trabalho foi criado com o intuito de relacionar os dois conceitos, a Engenharia Civil e *Data Mining.* A sociedade foca-se em encontrar padrões no seu dia-a-dia, é algo que instintivamente procuramos de forma a garantir alguma confiança perante um dado acontecimento. Se esse padrão se relevar forte o suficiente pode ainda ser capaz de alterar a forma de pensar de um individuo. Então o caminho que esta dissertação assume é no âmbito de encurtar a distância existente entre estes dois conceitos [1,2].

De muitas aplicações possíveis de *Data Mining* na da Engenharia Civil, uma das mais intuitivas será talvez o estudo da eficiência energética. Ao associar diferentes componentes construtivas com os gastos energéticos é de todo um bom partido para esta fase introdutória de DM na área. Sabendo que desde 2008 que o consumo energético de um país aumentou em cerca de 20 a 40% graças ao consumo energético torna-se um campo ótimo para exploração [3].

1.2. OBJETIVOS

O objetivo principal desta dissertação é a aplicação de uma metodologia de *Data Mining* à avaliação dos padrões de ambiente interior em habitações e da influência que as suas características construtivas têm na eficiência energética e conforto. Para poder alcançar este objetivo global definiram-se os seguintes objetivos parciais:

- Conhecer o conceito de *Data Mining* e os procedimentos de análise normalmente associados;
- Selecionar um caso de estudo adequado à aplicação de *Data Mining* que funcionasse como um exemplo de utilização dos procedimentos de análise;
- Avaliar a aplicabilidade de diferentes ferramentas de cálculo como suporte do processo de *Data Mining*;
- Concluir sobre as potencialidades da aplicação do processo de *Data Mining* e evidenciar as dificuldades presentes em cada passo do procedimento.

1.3. ORGANIZAÇÃO DO DOCUMENTO

Este documento se apresenta estruturado de uma forma cuidada para que cumpra da melhor forma os objetivos a que se propõe. Os seguintes capítulos estão organizados de modo a que o leitor consiga ter uma compreensão ordenada e estrutura do que é *Data Mining*, da sua aplicabilidade e quais os seus resultados.

No Capitulo 1, Introdução, é apresentado um enquadramento geral do tema desenvolvido bem como quais os objetivos a que esta dissertação se propõe a comprimir.

No Capítulo 2, Estado e Fundamentos da Arte, é introduzido ao leitor todo o processo de DM. É apresentada a definição de *Data Mining* e das diferentes fases que constituem este processo. Após a compreensão das diferentes fases optou-se por uma explicação sobre alguns fundamentos matemáticos por detrás da execução de alguns dos métodos utilizados. São ainda introduzias algumas das ferramentas disponíveis atualmente no mercado para a aplicação deste método. Por fim apresentam-se alguns trabalhos semelhantes já desenvolvidos na área.

O Capítulo 3, Metodologia e Bases de Dados, é responsável por demonstrar ao leitor qual a metodologia de *Data Mining* adotada ao longo deste trabalho. Com a metodologia descrita faltava especificar qual a base de dados a estudar e de que ela era composta. Em seguida apresenta-se uma comparação e escolha de entre as diferentes ferramentas disponíveis, demonstrando por fim a sua aplicabilidade.

A Análise de Resultados e Discussão é apresentada no Capítulo 4. Nele são apresentados os diferentes resultados obtidos ao longo da aplicação da metodologia proposta. Cada passo fornece um conjunto de soluções distinto, mas cada vez mais próximo de um objetivo comum, sendo essa a estruturação seguida pelo capítulo. É ainda apresentada a discussão e reflexão sobre estes mesmos resultados, de modo a fornecer uma melhor interpretação por parte do leitor.

O Capítulo 5, Conclusão, diz respeito às conclusões retiradas com a realização deste trabalho. Nele procura-se responder aos objetivos apresentados anteriormente. São ainda propostos alguns desenvolvimentos futuros, para que *Data Mining* possa estar cada vez mais ligado à Engenharia Civil.

2
ESTADO E FUNDAMENTOS DA ARTE

2.1. CONCEITOS

Segundo os autores Han and Kamber [1] existem duas abordagens diferentes para o conceito "*Data Mining"* (DM) uma em que corresponde ao processo a partir do qual é possível extrair de uma base de dados algum tipo de conhecimento e outra em que DM é apenas um passo num processo de extração de conhecimento de uma base dados. A primeira abordagem existe devido ao comodismo da sociedade que transformou um termo como "*Knowledge Discovery from Data*" (KDD), termo correto mas mais longo, num termo mais curto e fácil. A segunda abordagem reflete para o facto de que DM é apenas parte de um processo chamado "*Knowledge Discovery from Data"*. No presente trabalho considera – se *Data Mining* como uma analogia a KDD compreendendo também que DM representa uma fase de KDD.

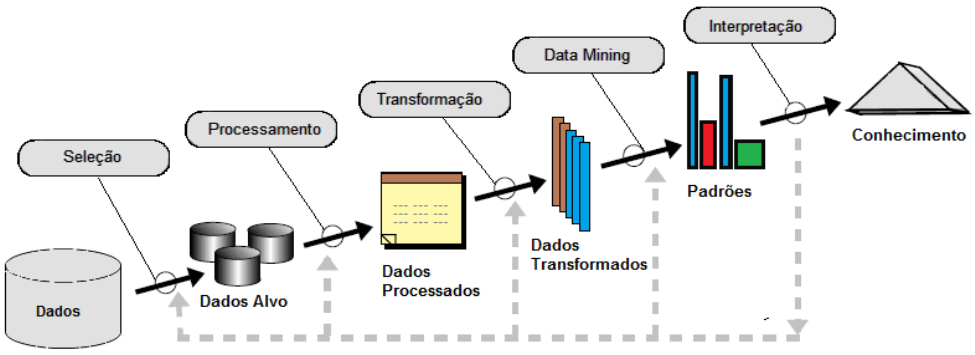


Fig. 2.1. Processo de *Data Mining* (adaptado de [1]).

Por definição, *Knowledge Discovery from Data* é um processo não trivial de identificação válida, fora do comum e potencialmente útil de compreensão de padrões específicos dos dados, definição segundo apresentam Usama, Gregory e Padhraic [4].

Sendo KDD um processo, este é então constituído por diversas fases, desde a recolha dos dados até ao conhecimento obtido como resultado final (Fig. 2.1). Esse processo é constituído por: Seleção dos dados, Processamento de dados, Transformação, Algoritmos de *Data mining* e por último Interpretação. Estes passos são de caracter cíclico, não estando vincados a apenas uma ocorrência. As metodologias

associadas a cada fase têm a particularidade de não serem fixas, isto é, podem ser utilizadas ao longo do processo de DM. Um exemplo é o método de formação de *clusters*, descrito mais à frente, a sua utilização pode ser feita em várias fases do processo.

2.1.1. SELEÇÃO DE DADOS

Começando por descrever os dados como um conjunto de medições recolhidas de um ambiente ou processo [2], estes são armazenados em estado "bruto" em diferentes tipos de bases de dados que carecem de trabalho para poderem ser interpretados. Esse trabalho vai permitir que a seleção de dados se possa realizar com vários níveis de detalhe, isto é, pode ser aplicada desde a escolha entre diferentes bases de dados até à seleção de um conjunto específico de dados [1].

Nesta primeira fase do processo de DM, podemos identificar que o nível mais geral corresponde à seleção das diferentes bases de dados disponíveis, bases de dados relacionais, armazéns de dados, bases de dados transacionais (armazenam transações) e sistemas de informação avançados. São exemplos: bases de dados temporais, espaciais, multimédia, a própria *World Wide Web*, entre outros. A nível mais específico encontramos pequenos conjuntos de dados dentro de uma base de dados [1].

Tendo em conta o âmbito deste trabalho, as bases de dados relacionais são as de maior interesse e correspondem a um conjunto de tabelas que contêm atributos específicos (colunas) que armazenam sequências de "n" elementos ordenados (tuplas) nas suas linhas, isto é, as tabelas representam cada objeto, identificado por um nome (atributo) específico e descrito por um conjunto de valores estando estas ligadas entre si (Fig. 2.2) [1].

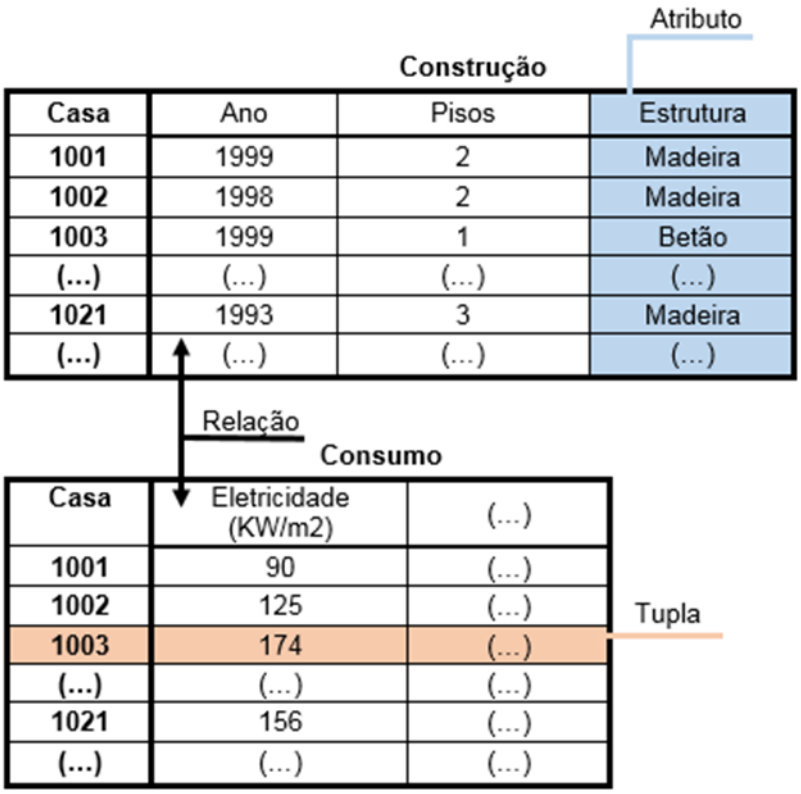


Fig. 2.2. Exemplificação de bases de dados relacionais.

A seleção de dados representa uma das etapas menos metódicas em que a sua execução depende da sensibilidade do investigador, do tipo e objetivo da investigação. Esta etapa consiste na seleção de um grupo de dados "alvo" provenientes de conjuntos de dados, subconjuntos de variáveis ou amostras de dados para o objetivo de análise pretendido [4].

2.1.2. PROCESSAMENTO DE DADOS

O processamento de dados tem como principal objetivo melhorar a eficiência do processo de DM, isto é, permitir trabalhar a maior quantidade de dados possível, durante o menor período possível e sem perder qualidade da informação. Este atributo surge devido ao facto de que a quantidade de informação armazenada é cada vez maior, aumentando o risco de falha do processo de armazenamento. Por outas palavras, os dados que pretendemos analisar através de técnicas de *data mining* estão incompletos (faltam valores de atributos ou certos atributos de interesse, ou podem conter apenas dados agregados), dispõe de ruído (contêm erros, ou *outliers*, valores que se desviam da normalidade), e inconsistência (exemplo, contendo discrepâncias nos códigos de departamento usados para categorizar um dado produto) [1].

Para colmatar tais problemas dispomos de métodos para limpeza e integração dos dados. Limpeza de dados é um processo que permite preencher valores em falta, suavizar o ruído identificando *outliers* e corrigindo dados inconsistentes. A integração dos dados diz respeito a um processo de diminuição do tempo e aumento da precisão do processo de DM. O facto de os dados serem provenientes de diferentes fontes ou bases de dados representa um problema que para além de interferir na coerência de informação reflete novamente problemas como o ruido e afins [1].

A questão da coerência remete para duas problemáticas, problemas de identidade (como é que dados diferentes, de bases de dados diferentes, podem-se relacionar) [1] e existência de redundância das variáveis que provoca um incremento no tamanho da base de dados aumentando o tempo de modelação através de ferramentas DM. Estas redundâncias estão então diretamente ligadas à correlação entre as variáveis sendo que podem ser tratadas com testes de correlação entre elas [2].

A falta de valores ocorre quando alguns atributos não foram introduzidos ou, por alguma razão, não foram registados durante o processo de registo (erros de equipamento). Existem duas abordagens principais para o problema, a mais intuitiva e fácil é ignorar esses valores. Contudo este processo só é aceitável se a quantidade de valores em falta for muito pequena relativamente ao dados em estudo. A abordagem mais trabalhosa, mas precisa, é preencher esses valores. A reposição de valores em falta pode ser executada à mão (valores fixos, médios ou reais), através de procedimentos de "*maximum likelihood*" ou imputação dos valores [1, 2].

Também o ruido de uma amostra é um fator a tratar nesta fase do processamento. A Fig. 2.3 apresenta um exemplo de ruido existente numa amostra de dados.

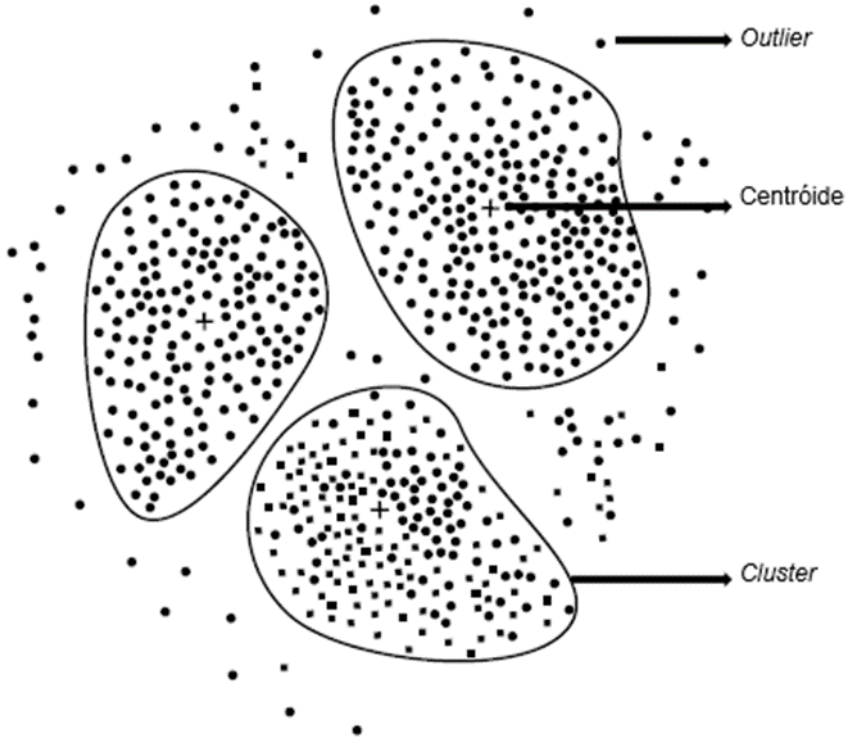


Fig. 2.3. Representação gráfica de um conjunto bidimensional de 3 *Clusters*, seus centróides e *Outliers* (adaptado [1]).

A remoção de ruído, ou suavização, tem como objetivo eliminar grandes disparidades nos valores existentes que podem afetar o desempenho de um processo de análise, principalmente em termos de classificação. Estes valores "ruído" podem criar pequenos *clusters* de um grupo de variáveis comprometendo a real aglutinação das mesmas no seu estudo. Como é possível verificar na Fig. 2.3, adaptada do livro de Han e Kamber, apresentam-se 3 *clusters*, delimitados pelas linhas, marcados com o símbolo "+" os seus centróides, representa o valor médio do cluster, e *outliers*, valores que ficam fora dos círculos e que correspondem a valores cuja sua ocorrência é muito rara com base nas observações restantes [1, 2, 5].

Em termos de problemas de identidade, estes apenas podem ser resolvidos através da refleção e organização das variáveis da melhor forma. Por fim a redundância pode ser levantada através de análises de correlação entre as variáveis, permitindo identificar aquelas que têm o mesmo significado perante a análise [1, 2].

2.1.3. TRANSFORMAÇÃO E REDUÇÃO DE DADOS

A transformação e redução de dados corresponde à última etapa de tratamento, antes da aplicação de ferramentas de *Data Mining*. Nesta fase são aplicados vários processos, processos de limpeza de dados (redução de ruído ou suavização), de redução de dados (agregação, seleção de conjuntos de atributos,…) e por fim de transformação (normalização), de modo a encontrar as melhores características para uma mais fácil análise tendo em conta o objetivo final [1, 4].

Da transformação resulta um conjunto de dados mais fácil de ler ou interpretar, contudo esse conjunto pode ainda ter dimensões significativas, sendo necessária a aplicação de uma redução da dimensão da base de dados. Esta redução tem como obrigação manter, o mais possível, a integridade da informação existente no conjunto de dados original [1].

Com a agregação é possível resumir os dados de que dispomos [1]. A Fig. 2.4 apresenta um exemplo do resultado deste processo, em que através dele é possível aglomerar as temperaturas horárias interiores de varias casas, em médias horárias da temperatura interior para cada casa.

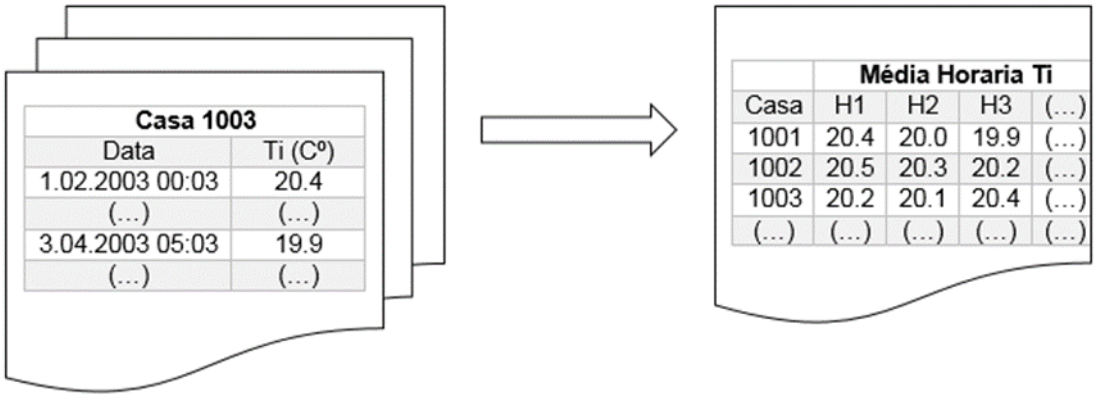


Fig. 2.4. Exemplo de agregação de dados (adaptado de [1]).

É possível selecionar subconjuntos de atributos com base em dados pouco ou nada relevantes e atributos redundantes determinados anteriormente diminuindo assim as dimensões da amostra em questão. Esta seleção traz um segundo benefício, que diz respeito à diminuição de variáveis, facilitando a futura análise de padrões. A melhor forma de selecionar este tipo de subconjuntos é determinando a significância das variáveis entre si [1].

Redução da dimensão é também um processo na redução de dados, em que existem mecanismos que são eficazes na redução das variáveis em análise, de que é exemplo a análise dos componentes principais explicada mais à frente.

A normalização é um processo em que os atributos são escalados para um específico intervalo (entre -1.0 e 1.0 ou entre 0.0 e 1.0), normalmente provenientes de processos de redução, tornando-se muito úteis em procedimentos de classificação, como *clusters*. [1, 2].

2.1.4. *DATA MINING* E A DESCOBERTA DE PADRÕES

Esta é a fase mais importante de todo este processo, sendo também responsável pela adaptação do nome de *Knowledge discovery from data* para *Data Mining.* Até agora todas as fases descritas correspondem à preparação dos dados. Foram descritos meios para que seja possível a aplicação desta fase da maneira mais eficiente. Como é possível observar na figura Fig. 2.1, esta fase tem como resultado a identificação de padrões [1, 2].

Um padrão consiste numa relação frequente num conjunto de dados. Essa associação que ocorre entre variáveis pode ser explicada pelo método da análise do cesto de compras (Fig. 2.5). Como é possível observar nessa figura, a compra de um computador está também associada à compra de um rato para

cada cliente, identificando-se assim um padrão. Como forma de representação desses padrões é possível identificar regras de associação entre as variáveis. No seguinte exemplo,

$$Computador\ (A) \rightarrow Rato(B)\ [suporte = 75\%, confiança = 100\%], \quad (2.1)$$

em grandes quantidades de dados, faz sentido avaliar o suporte e a confiança que dada regra contem. Suporte é visto como a frequência de ocorrência de um dado conjunto ($P(A \cup B)$) e confiança representa a probabilidade de ocorrência do acontecimento $A \cup B$ sabendo que A ocorre ($P(B|A)$).

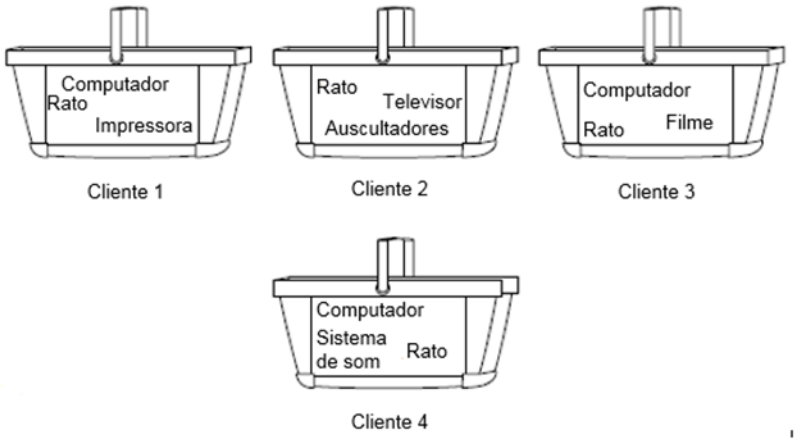


Fig. 2.5. Análise cesto de compras (adaptado de [1]).

As regras de associação são determinadas de diferentes modos, desde algoritmos específicos até conclusões com base na observação [1].

Determinados os padrões que um dado conjunto de dados possui é necessário identificá-los. Existem duas formas de classificar padrões encontrados, através de modelos de classificação e/ou de modelos de previsão [1].

Os modelos de classificação, em semelhança com o que aqui foi enunciado durante as fases de tratamento de dados, são responsáveis por prever rótulos para os diferentes casos observados. Estes modelos são criados segundo um processo de dois passos. No primeiro é selecionado um conjunto de dados teste e assumindo que cada tupla pertence a uma certa classe (é submetida a um rótulo). Este processo chama-se um processo de aprendizagem supervisionado, pois dispomos de informação sobre as classes a que cada tupla pertence. No caso de não conhecermos essa informação é necessária a aplicação de métodos, como os de clusterização para obtenção desses grupos (aprendizagem não supervisionada). O segundo passo consiste na aplicação do modelo aprendido para classificação [1, 2].

Os modelos de previsão são construídos de modo a prever uma função contínua e não apenas rótulos para um conjunto de dados. Estes modelos são constituídos por dois passos semelhantes aos apresentados para a classificação anterior. A diferença está no tipo de classificação, em que caiem os rótulos, visto que a classificação da tupla é feita de um modo contínuo e não categórico. As árvores de decisão são um bom exemplo deste tipo de modelos, contudo existem outros, como Classificação de *Naïve Bayesian*, *Support Vector Machines*, entre outros, para os quais que não houve oportunidade de os descrever e por isso aconselha-se a leitura da bibliografia indicada [1, 2].

2.1.5. INTERPRETAÇÃO DE RESULTADOS

A interpretação de resultados é um passo transversal a todas as análises efetuadas e não só aquelas que são baseadas em DM. Esta fase é a última fase deste processo e é nela que se retiram as conclusões. A visualização dos modelos/padrões obtidos tornam-se um fator determinante na obtenção de conclusões [4].

2.2. MÉTODOS DE ANÁLISE APLICÁVEIS

Como verificado anteriormente, *Data Mining* é um processo de organização, filtragem, compreensão e avaliação de dados, e para tal necessita de procedimentos e métodos matemáticos. Neste trabalho é assumido que o leitor tem compreensões básicas matemáticas e estatísticas.

Neste espaço que se segue são expostos alguns métodos estatísticos, usados em DM, bem como algumas noções que direta ou indiretamente são importantes para uma melhor compreensão dos mesmos.

2.2.1. ANÁLISE DE COMPONENTES PRINCIPAIS

A análise de componentes principais (ACP) é uma técnica de análise multivariada, análise simultânea de diversas variáveis, que analisa os dados selecionados em que as observações são constituídas por diversas variáveis dependentes, normalmente interligadas entre si. O seu objetivo é extrair um conjunto de variáveis ortogonais chamadas componentes principais [6].

No conceito aliado a *Data Mining*, ACP tenta encontrar para k variáveis ortogonais que melhor representam os dados em questão e em que $k \leq n$ (n=número de variáveis originais). Os atributos resultantes, componentes principais, são criados em ordem descendente de contribuição, sendo que o primeiro componente principal é aquele que apresenta uma maior porção da variância do conjunto de dados original [1].

O procedimento normal é considerar as primeiras componentes criadas que contenham entre 95% a 98% da variância dos dados originais. A ACP é aplicada normalmente seguindo um procedimento que se inicia por uma normalização das variáveis, equilibrando os intervalos entre atributos de modo a que os atributos de maior domínio não dominem os de menor domínio. Em seguida realiza-se o cálculo dos k vetores ortogonais para servirem como base dos dados normalizados, componentes principais com direção ortogonal entre si em que a base de dados original resulta da sua combinação linear destas. Encontrados os k vetores, estes são organizados por significância descendente, em que no fundo os componentes principais funcionam como um novo conjunto de eixos que fornecem importante informação sobre a sua variância. Por fim como os dados são organizados de forma descendente, é possível eliminar as componentes mais fracas, permitindo assim obter uma grande aproximação do conjunto de dados original com um conjunto de atributos mais pequeno [1, 2].

A Fig. 2.6 ilustra o processo da análise de componentes principais em que P1 representa a primeira componente principal e ortogonalmente a esta está representada P2, bem como a segunda componente principal.

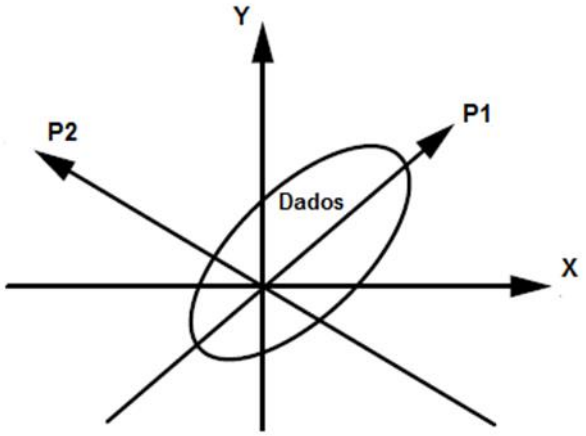


Fig. 2.6. Os dois primeiros componentes principais P1 e P2 (adaptado de [1]).

Considerando uma base de dados com n variáveis, $x_1, x_{2,} x_{3,\dots}, x_n$. A covariância entre as variáveis Xi e Xj para m objetos, pode ser estimada por:

$$s_{ij} = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x}_i)(x_{kj} - \overline{x}_j) \quad (2.2)$$

Notando que as covariâncias são simétricas, $s_{ij} = s_{ji}$, e que s_{ii} é a estimação da variância de X_i, s_i^2. A covariância está então relacionada com a correlação linear (entre x_i e $x_{j,}$ s_{ij}), sendo esta estimada por:

$$r_{ij}\frac{\sum_{k=1}^{n}(x_{ki}-\overline{x}_i)(x_{kj}-\overline{x}_j)}{(n-1)s_i s_j} = \frac{s_{ij}}{s_i s_j}, \text{ com } r_{ij} \in [-1,1] \quad (2.3)$$

É possível então verificar que a correlação pode ser interpretada como uma covariância normalizada.

Para obtenção das componentes principais do conjunto de dados em análise, é necessário encontrar combinações lineares não correlacionadas, cujas variâncias representadas sejam as maiores possíveis. Nesta linha de raciocínio entende-se então que a primeira componente principal corresponde à direção com máxima variância, a segunda componente principal constitui a direção seguinte não correlacionada que maximiza a variância que resta representar, e assim sucessivamente. Este processo de maximização é expresso pela equação:

$$(S - \lambda_i I)u_i = 0 \quad (2.4)$$

em que I representa matriz identidade, λ_i é um escalar e u_i é o vetor coluna, $m \times 1$, dos coeficientes de combinação linear. Os escalares λ_i são os valores próprios da matriz S e da resolução dos sistemas homogéneos (eq. 2.3). Para os diferentes valores próprios resultam os vetores próprios u_i (variáveis não correlacionadas). Normalmente, destes são selecionados os vetores que têm norma igual a 1 [7].

2.2.2. Análise de *Clusters*

A análise de *Clusters*, consiste no processo de agrupamento de objetos, criando grupos em que os diferentes objetos em estudo tenham grande semelhança. Um *cluster* diz respeito a um conjunto de dados que é similar entre si ficando por isso num certo grupo durante o processo de formação de clusters, (clusterização). Esta análise pode ter diferentes objetivos, sendo os principais: Identificação de *outliers*, para uma suavização dos dados ou estes podem ser o foco de estudo, e segmentação dos dados para uma divisão em subconjuntos [1].

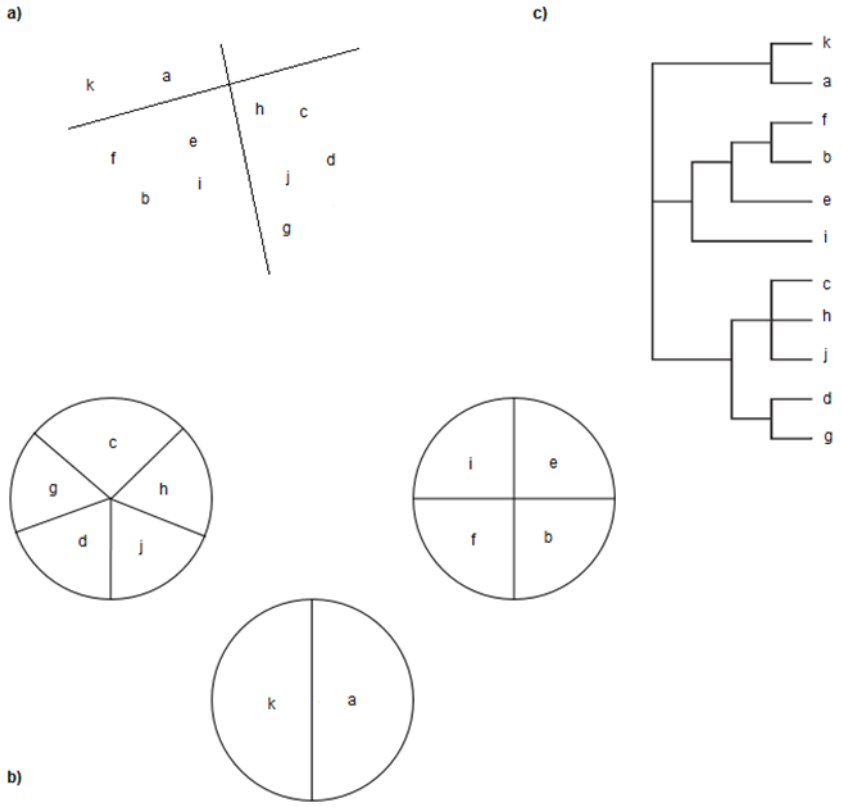


Fig. 2.7. Diferentes representações de clusters. Em a) e b) representações os objetos são representados e distribuidos pelo espaço mostrando cada cluster e em c) representação hierárquica, dendrograma (adaptado de [1]).

Este processo tem por base a aprendizagem através da observação e os seus resultados têm a forma de diferentes diagramas que demonstram como os objetos se distribuem pelos diferentes grupos, tal como é possível observar nas Fig. 2.3 e Fig. 2.7 [1, 8].

Existem vários métodos para a análise de *clusters*: métodos hierárquicos, Ward's e *complete linkage*, entre outros, e métodos não hierárquicos, como é exemplo o algoritmo de divisão *K-means*. Apenas se apresentam aqueles os mais relevantes para este trabalho sendo que para uma compreensão mais completa sobre o assunto aconselha-se a leitura da bibliografia indicada [1, 8].

Os métodos hierárquicos dividem-se em métodos aglomerativos e divisivos. Os aglomerativos iniciam o processo agrupando cada elemento no seu próprio *cluster*, continuando a agrupar os diferentes *clusters* até chegar a um único *cluster* contendo todos os objetos. Os métodos divisivos, partem da singularidade, um *cluster* contendo todos os objetos, que se vai subdividindo em *clusters* mais pequenos até existir um *cluster* para cada objeto (Fig. 2.8) [1, 8].

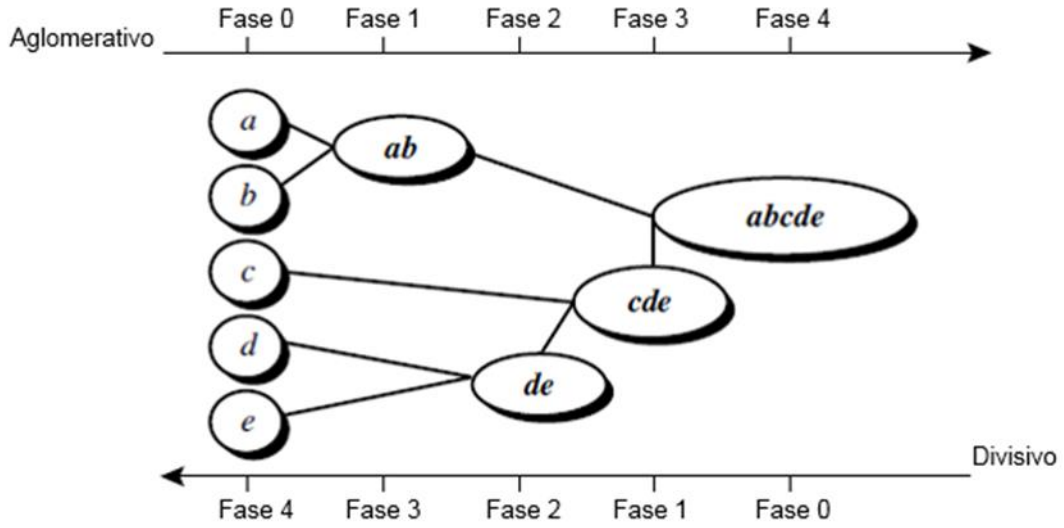


Fig. 2.8. Ilustração dos métodos aglomerativos e divisivos dos objetos $\{a, b, c, d, e\}$ (adaptado de [1]).

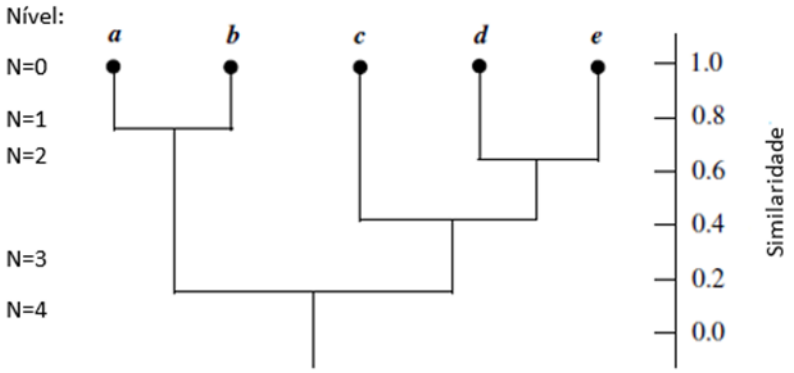


Fig. 2.9. Dendrograma

O resultado deste processo pode ser visualizado através de um dendrograma. A Fig. 2.9 apresenta o dendrograma resultado dos 5 objetos presentes na Fig. 2.8 [1].

A realização desta agregação de objetos depende da condição de similaridade, ou distância métrica entre eles, sendo esta a componente fundamental na formação dos *clusters*. De entre as variadas formas de calcular essa distância, aquelas que representam importância para este trabalho, considerando que a distância entre dois objetos, p e p', é representada por $|p - p'|$; os *clusters* são representados por C e o número de objetos num dado cluster por n, são [1]:

$$d_{max}(C_i, C_j) = max_{p \in C_i, p' \in C_j} |p - p'|, \tag{2.5}$$

em que d_{max} representa a distância máxima. Este processo termina com a distância máxima entre os *clusters* mais próximos a exceder um limite arbitrário. Este tipo de algoritmos tende a minimizar o aumento do raio do *cluster* a cada iteração, produzindo clusters de alta qualidade quando a informação se encontra mais compactada mas também muito sensíveis aos *outliers* [1, 8];

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|, \quad (2.6)$$

em que d_{avg} representa a distância média. Os algoritmos de formação destes *clusters* que utilizam este cálculo de distância têm como objetivo evitar que estes não sejam tão sensíveis a *outliers* [8].

Descrevendo agora o mais importante método de clusterização não hierárquico, temos o algoritmo *k-means*. Aquando o uso deste a similaridade é calculada através da distância euclidiana, D (Fig. 2.10). Por definição:

$$D(C_i, C_j) = \sqrt{\sum_{i=1}^{n} (p_i - p'_i)^2} \quad (2.7)$$

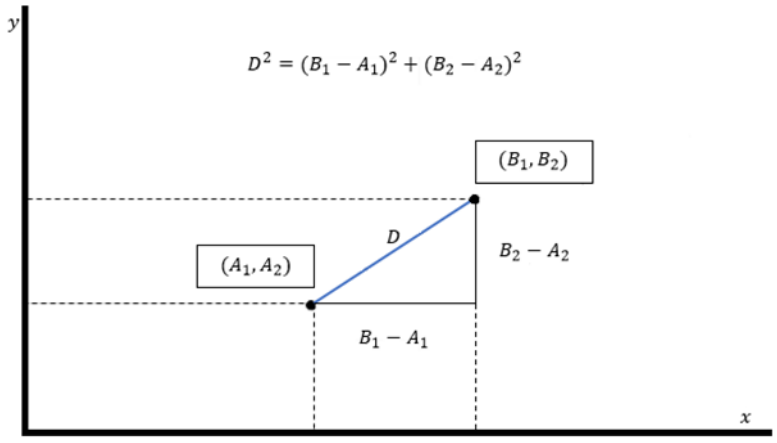


Fig. 2.10. Exemplificação do cálculo da distância euclidiana.

O algoritmo "*k-means"*, é um método baseado em centroides (Fig. 2.3) que agrupa um conjunto de n objetos em k *clusters*. Este método apresenta algumas desvantagens, tais como a determinação do parâmetro k, é um método muito sensível ao ruido existente nos dados e a *outliers*, principalmente porque um pequeno conjunto de dados é suficiente para influenciar a média calculada [1].

Comparando então a aplicação dos dois métodos, pode-se concluir que é possível maximizar a utilização de cada um deles efetuando primeiro uma análise hierárquica para definir o número de *clusters* e comparar essa análise com o uso de um método não hierárquico, usando o número de *clusters* anteriormente obtido [1, 8].

2.2.3. Árvores de decisão

As árvores de decisão são um método de classificação, correspondente a um modelo de previsão. Este diagrama tipo fluxograma (Fig. 2.11) é composto por quatro componentes diferentes: os nós internos, os ramos, os nós terminais (ou nós folha) e o nó raiz, no topo.

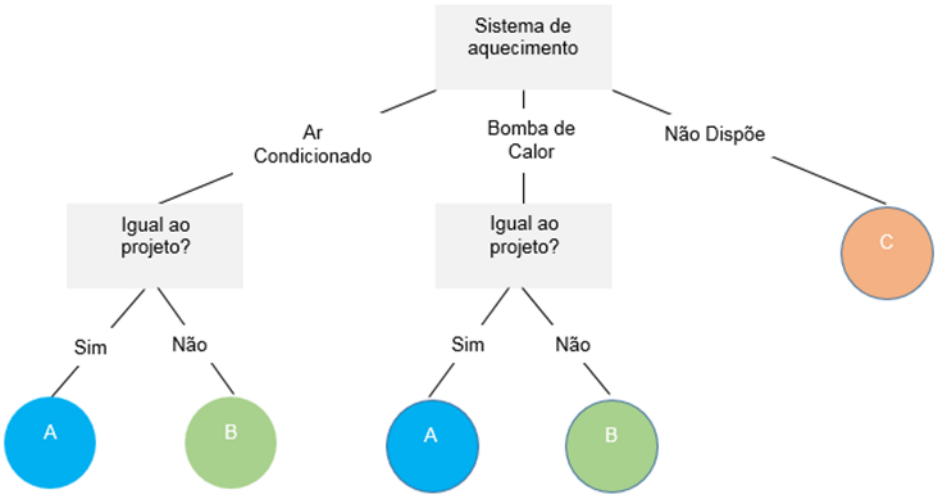


Fig. 2.11. Exemplo de uma classificação tipo árvore de decisão (adaptado de [1]).

Os nós internos dizem respeito a testes de atributos, em que cada ramo representa um resultado desse teste. Por fim, os nós terminais asseguram a classificação. Dada uma tupla específica, cuja classe de associação é desconhecida, os valores dos atributos desse grupo são testados à medida que vão percorrendo a árvore. Após identificado o caminho desde a raiz até ao nó terminal fica definida a classificação desse atributo, ficando assim criado o modelo de previsão [1].

Este método apresenta grande sucesso porque não necessita de grande preparação para a sua execução. Não carece, por exemplo, da necessidade de definir parâmetros específicos. Uma outra razão para a sua preferência é o fato da sua representação ser bastante intuitiva, tornando a aprendizagem e classificação dos atributos muito rápida e eficaz [1].

2.3. PROGRAMAS DE CÁLCULO

Os programas de cálculo disponíveis atualmente no mercado para proceder a análise de *Data Mining* dividem-se em dois grupos, aqueles que permitem armazenamento da informação (criação de bases de dados) e aqueles que têm como objetivo a realização do trabalho estatístico necessário ao processo de "*Data Mining"*.

2.3.1. ARMAZENAMENTO DE DADOS

A existência de um estudo sobre um conjunto de dados implica o seu armazenamento. Esse armazenamento é necessário ao longo das várias fases de qualquer processo analítico e dá origem às bases de dados.

Na atualidade existem várias ferramentas, gratuitas ou não, disponíveis para a criação de bases de dados. Neste texto apenas são referenciadas algumas delas, tais como, Microsoft Office Excel, Microsoft Office Access e Oracle *Database* [9, 10, 11].

Mais a frente será elaborada uma comparação entre este *softwares*. Essa análise terá como objetivo identificar as várias diferenças entre eles e escolher aquele que mais se adapta a este trabalho.

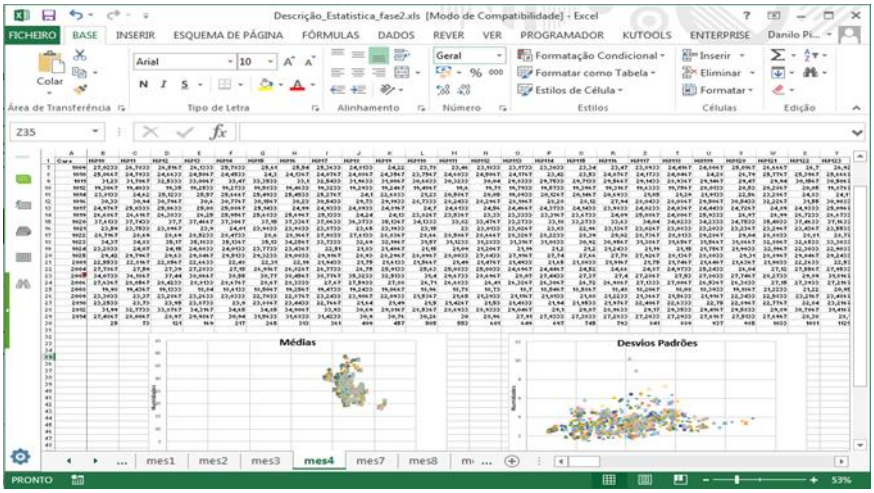


Fig. 2.12. Representação de uma folha de cálculo de Excel em que se podem verificar as suas aplicações distintas (base de dados e ferramenta de cálculo).

De todos os programas aqui referidos, são identificáveis dois grupos, Microsoft Office Access e Oracle *Database* fazem parte de um grupo e Microsoft Office Excel corresponde a outro. Esta diferenciação existe porque o Excel corresponde a um programa de cálculo, sendo adaptável para funcionar como base de dados, isto é, permitir aplicar funções de organização e filtragem dos dados. A Fig. 2.12 comprova este fato pois é possível observar na parte superior da folha de cálculo os diferentes atributos e tuplas e na parte inferior dois gráficos de dispersão obtidos através de cálculo.

2.3.2. PROGRAMAS DE ANÁLISE AVANÇADA

As ferramentas de análise avançada são as que permitem que todo este estudo seja possível. Têm como objetivo a implementação dos algoritmos estatísticos e matemáticos de modo a criar condições para a interpretação dos dados. Em comparação com as ferramentas de armazenamento, estas apresentam um âmbito mais vasto, em que é possível idealizar uma divisão destes programas, os de análise estatística e os específicos de *Data Mining*.

As ferramentas que aqui se consideraram como ferramentas de análise estatística, são aquelas cuja função principal é a aplicação de métodos estatísticos que podem ser aplicados no processo de DM. As restantes são consideradas de "caixas negras" (*Black-Box*) porque fornecem ferramentas aplicáveis neste processo sem conhecimento do como ou quais os parâmetros (alguns) que estes *softwares* usam no processamento de dados. Esta problemática será abordada mais à frente com exemplos que a ilustram melhor.

Das várias ferramentas de análise estatística existentes, as de salientar são SPSS, STATISTICA, RStudio e MATLAB. Estas duas últimas são ferramentas de cálculo matemático que permitem a aplicação de análises estatísticas, enquanto as primeiras são ferramentas que se focam na análise estatística em si.

SPSS e STATISTICA são dois programas de análise estatística não gratuitos. Ambos dispõem de ferramentas de cálculo estatístico, permitindo a análise, gestão, visualização e integração de processos mais complexos como os de DM. Ambos os *softwares* se focam na acessibilidade e facilidade permitindo um cálculo estatístico intuitivo com base numa folha de cálculo. São ambos muito semelhantes, tanto na interface como na metodologia de trabalho como será possível verificar no Capítulo 3. Apenas por curiosidade o programa SPSS foi inicialmente criado em 1968 por três jovens (Norman Nie, Hadlai Hull, and Dale Bent) com a visão de incorporar os resultados estatísticos na tomada de decisão e apenas

em 2009 foi comprado pela empresa IBM. Este fato é interessante porque um dos objetivos do processo de DM é o apoio na tomada de decisão, não ficando esta apenas baseada na intuição [7, 12, 13, 14].

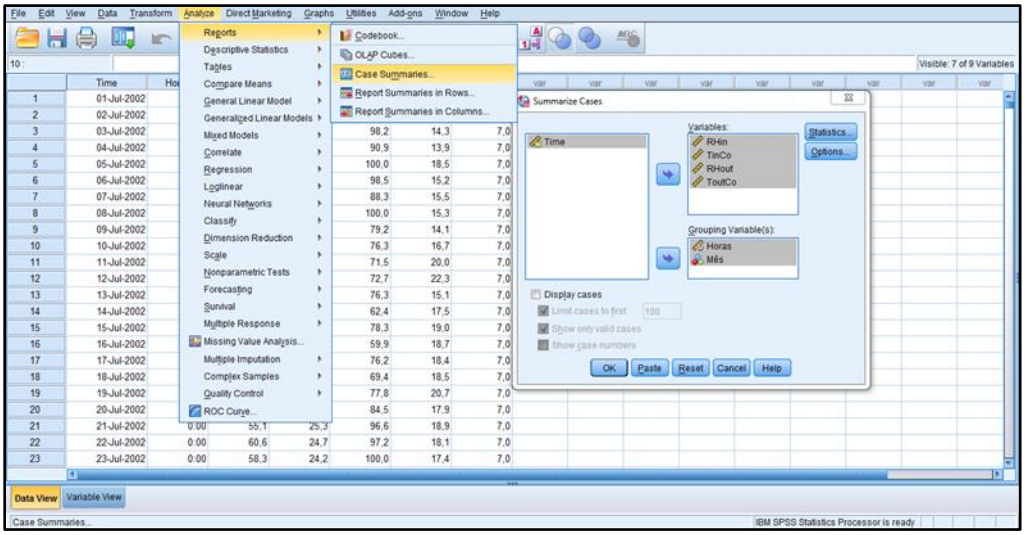


Fig. 2.13. Exemplo da interface de SPSS.

MATLAB e Rstudio, dois programas semelhantes na sua forma de trabalhar, apostam na liberdade de cálculo. MATLAB pertence à empresa *The MathWorks* e tem como função principal ser um programa de cálculo matemático. RStudio, *software* gratuito, pertence à *R Development Core team* e tem como base a linguagem R (*Language and Environment for Statistical Computing*). Como referido, estes programas apresentam liberdade dentro do campo de análise, pois permitem a introdução de algoritmos bastante específicos capazes de fornecerem a análise mais detalhada que podemos encontrar no mercado. Contudo esta liberdade tem um preço que é a necessidade de aprender estes dois tipos de linguagem que diferem do *visual basic* atual e aproximam-se da linguagem de comandos [7, 15, 16].

Por fim apresentam se programas agrupados como caixas negras, de que são exemplos o Weka, desenvolvido pela universidade de Waikato (Nova Zelândia), e rapidminer. Estes dois *softwares*, gratuitos, são caracterizados por conterem algoritmos específicos para a execução de *data mining*. A sua utilização baseia-se sobretudo na aplicação direta desses algoritmos [17, 18].

2.4. *DATA MINING* E O COMPORTAMENTO DOS EDIFÍCIOS

Data Mining é um conceito que já existe há algum tempo mas ligado sobretudo às áreas da gestão, economia, sociologia, entre outros [1, 2, 3]. Contudo já existem alguns trabalhos que tendem a aproximar a construção do processo de DM. Os trabalhos existentes remetem na sua maioria para conceitos como comportamento dos ocupantes, deteção de falhas e eficiência energética [3, 19, 20, 21, 22, 23, 24].

A questão da aplicação de técnicas de DM a conceitos ligados a construção civil é de natureza recente e está presente em alguns dos estudos que aqui se apresentam. Na análise que se segue são enunciados alguns estudos que tiveram grande influência na realização deste trabalho bem como outros que representam estudos semelhantes que indicam o progresso de que esta metodologia tem vindo a ser submetida no ramo da construção.

2.4.1. *DATA MINING* E A EFICIÊNCIA ENERGÉTICA.

Um dos estudos que teve grande impacto na resolução deste trabalho foi o que Xiaoxin Ren, Da Yan e Tianzhen Hong realizaram, "*Data mining of space heating system performance in affordable housing*". O estudo tem como objetivo o uso de técnicas de DM para análise da eficiência de aquecimento em casas de preços acessíveis, promovendo informações sobre qual a necessidade de aquecimento deste tipo de habitações, se existem padrões únicos nos perfis de temperatura, a existência de alguma ligação entre os perfis de temperatura das habitações, o sistema em uso e a necessidade energética para o aquecimento e quais as estratégias que podem ser adotadas para melhorar o funcionamento do aquecimento [3].

Um dos pontos de grande interesse deste artigo foi a sua metodologia, semelhante aquela que é introduzida no início deste capítulo (Fig. 2.1), provando que esta é passível de ser aplicada no estudo de conceitos ligados aos edifícios. *Data Mining* foi usado com dois propósitos, a realização de uma análise descritiva e de previsão. A primeira diz respeito à utilização de ferramentas como análise de *clusters* e as regras de associação para uma perceção do tipo de padrões existentes na necessidade de temperatura destes apartamentos. A análise de previsão teve como principal componente a criação de classificações para a descoberta de um modelo de uma determinada classe em função de outros atributos como explicado mais a frente. Uma das formas de demonstração dos resultados por parte destes autores foi a aplicação das árvores de decisão introduzidas anteriormente.

Esta análise foi realizada num conjunto de 62 apartamentos com medições de temperatura interiores com um intervalo de 10 minutos durante 3 meses. Foi também contabilizado o consumo de gás e o tempo em que a lareira estaria desligada para futura comparação com o consumo elétrico. Foram criados perfis de temperatura para todos os andares do edifício englobando os 62 apartamentos, e através deles obteve – se a sua divisão em *clusters*. Estes grupos foram formados segundo um conjunto de 24x4 variáveis, isto é, usaram a média da temperatura horária e desvio padrão desde as 00 até as 23, separando os fins-de-semana dos dias úteis.

Após a clusterização os autores estudaram a variação temporária do estado ON/OFF dos termostatos. Desta análise consideraram uma classificação baseada nos rácios de ON/OFF. Semelhante processo foi elaborado e foram criadas classificações no consumo de gás diário.

Com a obtenção destes tipos de classificações foi realizada a análise de previsão que deu origem a uma árvore de decisão capaz de demonstrar intuitivamente os resultados obtidos.

Este estudo foi capaz de concluir que efetivamente *data mining* é uma ferramenta que tem muito potencial na avaliação de desempenhos energéticos de construções, permitindo desvendar padrões que refletem o consumo relativamente ao nível de conforto térmico exigido. Este estudo refletiu que métodos como os da formação de *clusters* e árvores de decisão para propósitos de classificação podem ser bastante uteis na obtenção de informação mais intuitiva. Foi ainda possível relacionar o uso do termostato com o consumo atingir um nível de conforto térmico requerido, constituindo este o passo anterior a criação de rotinas para um melhoramento da eficiência energética.

Um outro artigo de bastante interesse, "*Mining building performace data for energy-efficient operation",* apresenta um trabalho em que são comparados três modelos de classificação distintos, Naïve Byes, árvores de decisão e *support vector machine* (SVM). Este estudo aplicou estes três modelos de classificação a um modelo térmico, de modo a determinar o conforto térmico dos ocupantes (usando como variáveis, temperatura interior, humidade relativa, velocidade do ar, entre outras), e um modelo de iluminação [19].

Este estudo baseia-se no estudo de 70 quartos, de baixo consumo, num edifício irlandês. Este edifício dispunha de captação de energia solar, bombas de calor e sistemas de recuperação de calor. Os valores obtidos ditaram bons resultados para o processo de DM aplicado ao estudo do desempenho dos edifícios, relativamente ao conforto térmico dos ocupantes, à iluminação dos quartos e ao baixo consumo energético dos quartos. Foi possível concluir que dos modelos de previsão adotados, os de melhor comportamento foram Naïve Byes, seguido-se as árvores de decisão na identificação do uso dos quartos, enquanto o SVM teve um desempenho pobre. O método das árvores de decisão, tive um melhor desempenho na identificação dos quartos com melhor luminosidade, e Naïve Byes apresentou o melhor resultado na estimativa do conforto térmico [19].

Existem ainda outros estudos que demonstraram bastante foco na averiguação sobre os modelos de classificação de padrões utilizados. [20, 21, 22, 23, 24] A grande diferença do método anterior é que os modelos de classificação utilizados tem por base as regras de associação, mais precisamente a comparação entre as regras de associação, demonstrando a capacidade destas desempenharem um bom papel em processos de KDD. À margem da especificidade de cada modelo estudado, estas referências bibliográficas concluem que DM é um processo com futuro nesta área. Para além destes estudos que se baseiam na estrutura até agora apresentada, existem outros que dizem respeito á mesma metodologia e objetivos mas relativos á fase de projeto [25].

2.4.2. PARA ALÉM DE *DATA MINING*

Outros estudos apresentam caracter mais avançado, sendo realizados com objetivos de obterem resultados em "*machine learning*" e a previsão de consumos. Este tipo de artigos apresentam carácter mais avançado e um pouco distantes daquilo que DM pretende, contudo o uso de algoritmos e ferramentas é semelhante, o que leva a que sejam aqui também indicados. Os processos de *"machine learning"* são processos que dizem respeito a automação, a uma aprendizagem automática por parte da máquina para que esta seja capaz de aperfeiçoar o seu desempenho. Este tema já cai fora do âmbito deste trabalho mas, visto que DM é um dos processos que está incubado no seio deste tipo de aprendizagem (estes processos necessitam da análise de grandes quantidades de dados, formação de padrões e por fim "aprenderem" com eles), apresentam – se alguns estudos na área [26, 27, 28, 29].

3
METODOLOGIA E BASE DE DADOS

3.1. METODOLOGIA

No presente capítulo é apresentada a metodologia utilizada para a identificação dos perfis energéticos de um conjunto de casas através da aplicação de técnicas de *data Mining*. Como apresentado anteriormente DM corresponde a um processo longo e muito diversificado. A criação de perfis ou padrões têm no seu percurso uma vasta possibilidade de ferramentas e aplicações.

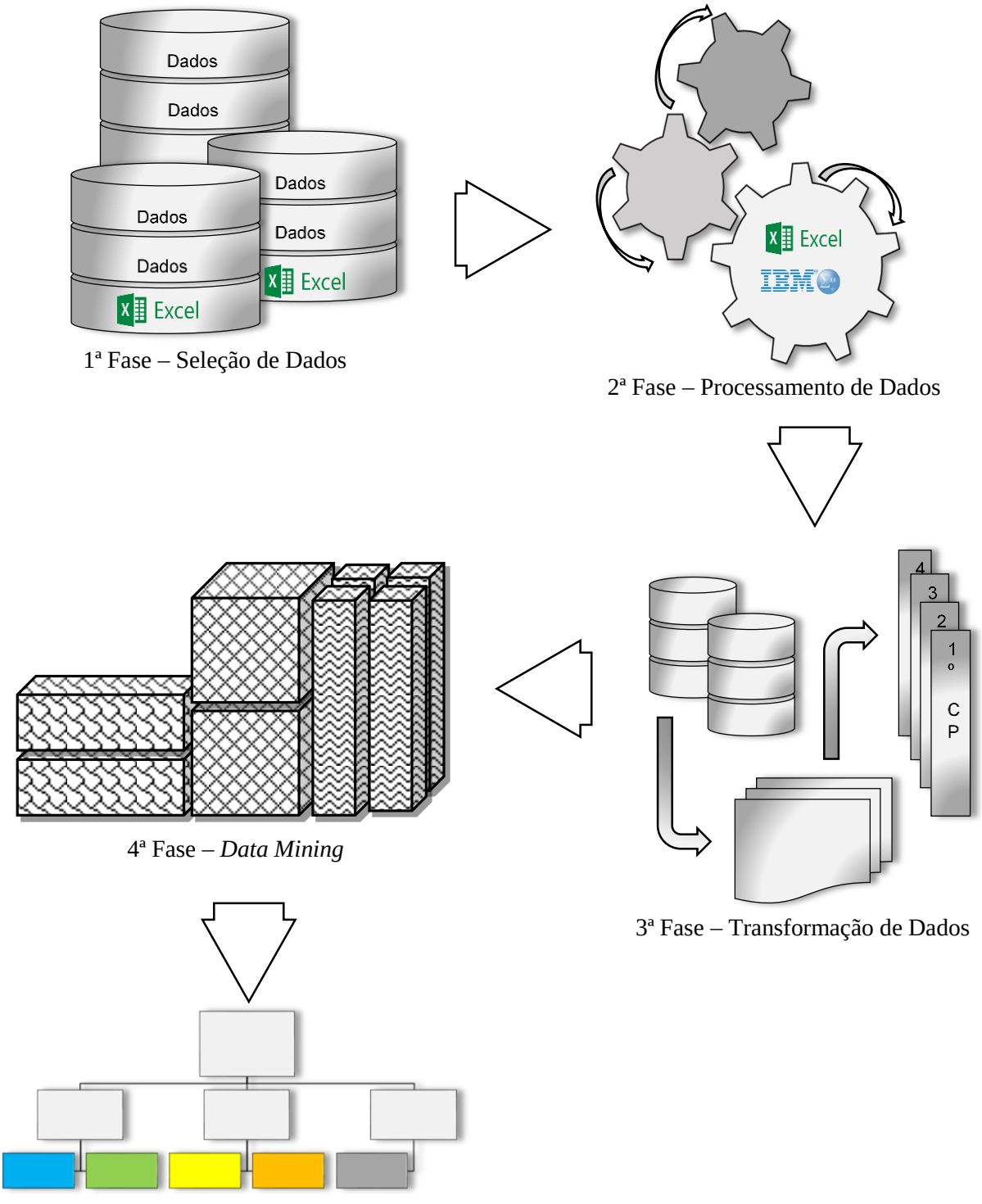
Data Mining é um processo estruturado que dispões de várias fases (Fig. 2.1). A sua estrutura não é muito suscetível de ser alterada, contudo dentro de cada etapa os métodos que são utilizados dependem de utilizador e dos dados de que dispõe.

A forma sequencial em como a aplicação do processo de *Data Mining* foi realizado ao longo desta dissertação é o que se apresenta na Fig. 3.1. Em seguida apresenta-se uma breve descrição sobre cada um dos passos apresentados nessa figura:

1. A primeira fase deste processo, Seleção de Dados, é responsável por selecionar quais os dados a estudar tendo em conta o objetivo do estudo. A seleção aqui presente tem como base o conforto térmico e a eficiência energética. Para que a escolha do conjunto de dados seja a mais rentável possível a base de dados é estudada e confrontada com o objetivo proposto;
2. Processamento de Dados, corresponde à segunda fase da metodologia. É responsável por processar os dados de modo a serem preparados para análise. Para tal, são utilizados três processos principais: Remoção de *outliers*, ou suavização dos dados; Uniformização dos dados, definindo um período de análise fixo bem como uma formatação fácil de trabalhar e compatível com o *software* escolhido; Reflexão sobre os dados em falta, o que implica a remoção ou reposição dos valores omissos, com base na quantidade de informação inexistente;
3. A terceira fase diz respeito à Transformação de Dados e que corresponde à última fase de preparação da informação. Esta fase é constituída por processo de agregação, redução e normalização. Os processos de agregação e redução correspondem a aglomeração de conjuntos de dados e à redução do número de variáveis presente. Para tal é adotado um estudo sobre a correlação entre as variáveis através do cálculo dos coeficientes de correlação. No caso de as matrizes de correlação apresentarem resultados fracos, passa-se à aplicação da análise dos componentes principais "forçando" a redução. O processo de normalização escala a grandeza das variáveis e será realizada em simultâneo com os processos de redução;
4. *Data Mining*, a quarta e mais icónica fase deste método. Nela são identificados padrões presentes no conjunto de dados transformados. Estes padrões são encontrados com o intuito de classificados. A classificação atribuída corresponde à grande objetivo deste processo. O

reconhecimento de padrões irá ser feito com base na análise de *clusters*, aplicando primeiro métodos hierárquicos e posteriormente o algoritmo *k-means*. A classificação será do tipo contínuo, usando o modelo de classificação Árvore de Decisão;

5. A última fase desta metodologia corresponde à Interpretação de Resultados. Nesta fase são analisadas as classificações obtidas anteriormente e retiradas conclusões.



1ª Fase – Seleção de Dados

2ª Fase – Processamento de Dados

3ª Fase – Transformação de Dados

4ª Fase – *Data Mining*

5ª Fase – Interpretação de Resultados

Fig. 3.1. A metodologia.

3.2. DESCRIÇÃO DA BASE DE DADOS

A base de dados neste estudo foi fornecida pelo Annex 55 *Reliability of Energy Efficient Building Retrofitting- Probability Assessment of Performance and Cost* realizado pelos Autores Nuno M. M. Ramos e John Grunewald. O documento é resultado do programa *Energy in Buildings and communities* (EBC) da *International Energy Agency* cujo objetivo é desenvolver e facilitar a integração de novos processos e ou tecnologias em áreas como a eficiência energética ou edifícios e comunidades sustentáveis, entre outros [30].

Os dados tiveram origem em estudos realizados pelas universidades tecnológicas de Tampere e Helsinki, Finlândia, com os responsáveis Elina Manelius e Juha Vinha. A recolha de dados consistiu em guardar informações sobre a temperatura interior, características de ventilação e de estanquidade em dois grupos diferentes de edifícios. O primeiro realizado entre 2002 e 2004 para 102 casas com estrutura de madeira. O segundo grupo, realizado entre 2005 e 2008, foi constituído por 50 casas com estrutura em betão, 20 casas de madeira tipo tronco e 56 apartamentos. Em acréscimo às características medidas foram também conhecidas informações como o tipo de ventilação, idade, método de construção, consumos energéticos, sistemas de aquecimento, entre outros parâmetros [30].

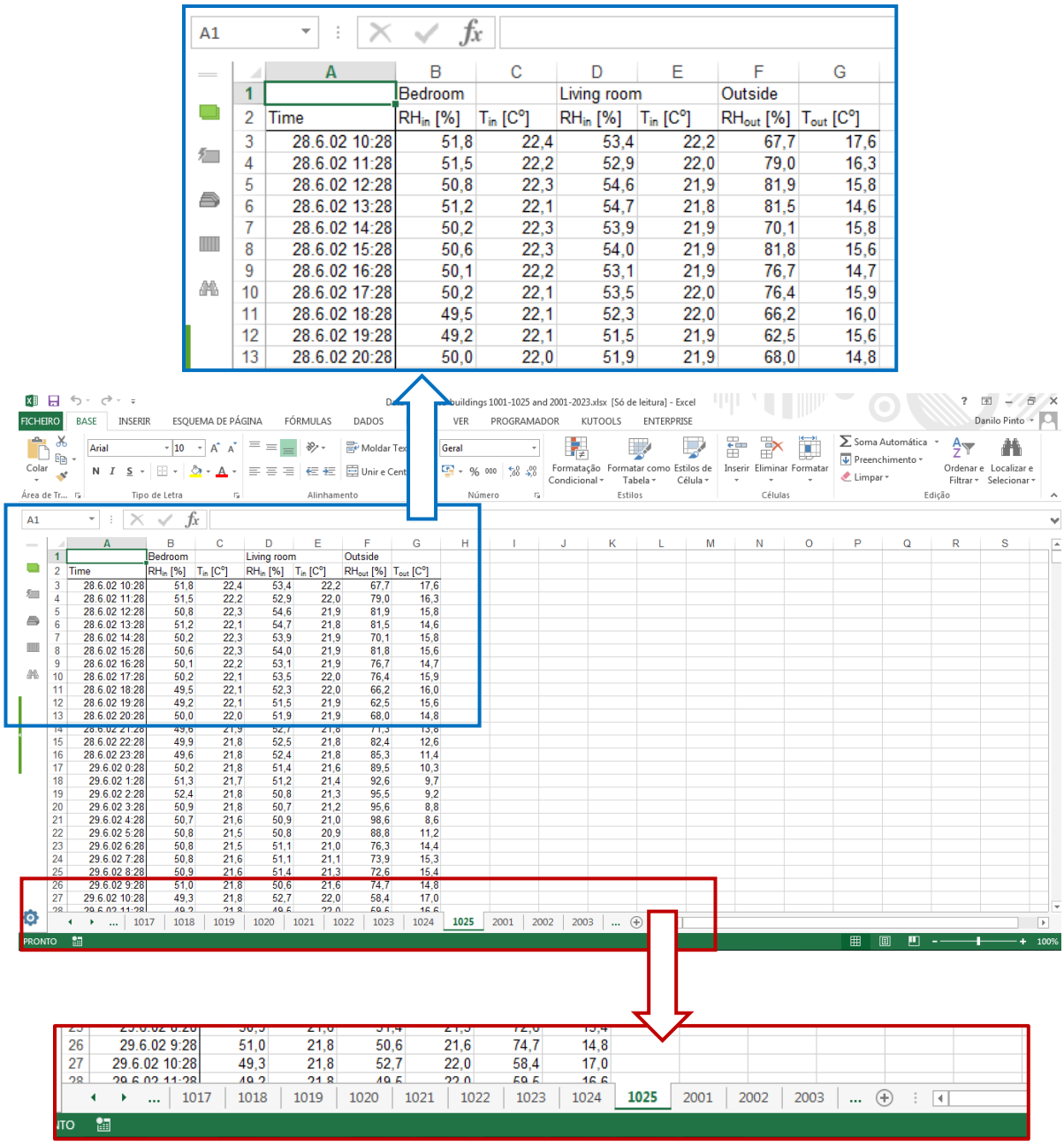
3.2.1. MORADIAS DE CONSTRUÇÃO LEVE

Este tipo de habitações individuais, com estrutura e placas de revestimento exterior em madeira, fazem parte do primeiro grupo enunciado anteriormente. A Fig. 3.2 permite uma visualização do tipo de habitações estudadas [30].

Fig. 3.2. Casas com estrutura e revestimento em madeira.
À direita o pormenor construtivo da fachada (retirado de [30]).

Nestes casos as medições de temperatura e humidade relativa foram realizadas entre Junho de 2002 e Junho de 2003 para a primeira metade das habitações e de Junho de 2003 a Junho de 2004 para as restantes. Estes parâmetros foram registados pelos aparelhos a cada hora durante períodos consecutivos de um ano. Na normalidade existiam apenas dois aparelhos de medição por habitação, um numa das paredes divisórias do quarto principal e outro no exterior da habitação, em locais em que a chuva e o sol não tivessem muita influência nas medições. Por vezes, algumas casas dispunham de um outro aparelho interior na sala de estar. Durante o segundo ano, as medições exteriores foram condicionadas na ocorrência de uma anomalia o que levou a não existência de dados exteriores nesse período de análise

[30]. A Fig. 3.3. apresenta uma folha de cálculo Excel em que foram armazenados os dados recolhidos pelos aparelhos.



		Bedroom		Living room		Outside	
	Time	RH_{in} [%]	T_{in} [C°]	RH_{in} [%]	T_{in} [C°]	RH_{out} [%]	T_{out} [C°]
3	28.6.02 10:28	51,8	22,4	53,4	22,2	67,7	17,6
4	28.6.02 11:28	51,5	22,2	52,9	22,0	79,0	16,3
5	28.6.02 12:28	50,8	22,3	54,6	21,9	81,9	15,8
6	28.6.02 13:28	51,2	22,1	54,7	21,8	81,5	14,6
7	28.6.02 14:28	50,2	22,3	53,9	21,9	70,1	15,8
8	28.6.02 15:28	50,6	22,3	54,0	21,9	81,8	15,6
9	28.6.02 16:28	50,1	22,2	53,1	21,9	76,7	14,7
10	28.6.02 17:28	50,2	22,1	53,5	22,0	76,4	15,9
11	28.6.02 18:28	49,5	22,1	52,3	22,0	66,2	16,0
12	28.6.02 19:28	49,2	22,1	51,5	21,9	62,5	15,6
13	28.6.02 20:28	50,0	22,0	51,9	21,9	68,0	14,8

26	29.6.02 9:28	51,0	21,8	50,6	21,6	74,7	14,8
27	29.6.02 10:28	49,3	21,8	52,7	22,0	58,4	17,0

1017 1018 1019 1020 1021 1022 1023 1024 **1025** 2001 2002 2003

Fig. 3.3. Folhas de Registo das Medições efetuadas. Em cima pode ver-se uma secção das medições e a sua organização. Em baixo é possível observar o código associado a cada habitação.

Como é possível observar as casas são identificadas com um código característico e foram armazenados medições horárias, referentes a temperatura e humidade do quarto principal, sala e exterior da habitação. A Fig. 3.3 diz respeito ao primeiro ano de medições e abrange as habitações 1001 a 1025 e 2001 a 2023. A informação relativa as restantes habitações e ao ano seguinte estão armazenadas noutro ficheiro de igual configuração [30].

A permeabilidade do ar foi recolhida através de testes de pressurização e despressurização e a ventilação foi estudada através da medição dos caudais de ar (admissão e extração) e das taxas de renovações horárias. Foram ainda recolhidas informações relativas ao número de moradores, características construtivas dos edifícios, sistemas de aquecimento e ventilação e consumos energéticos [30]. A Fig. 3.4. apresenta a base de dados em que foram registadas as informações sobre as características da habitação. Algumas dessas características foram apresentadas no parágrafo anterior.

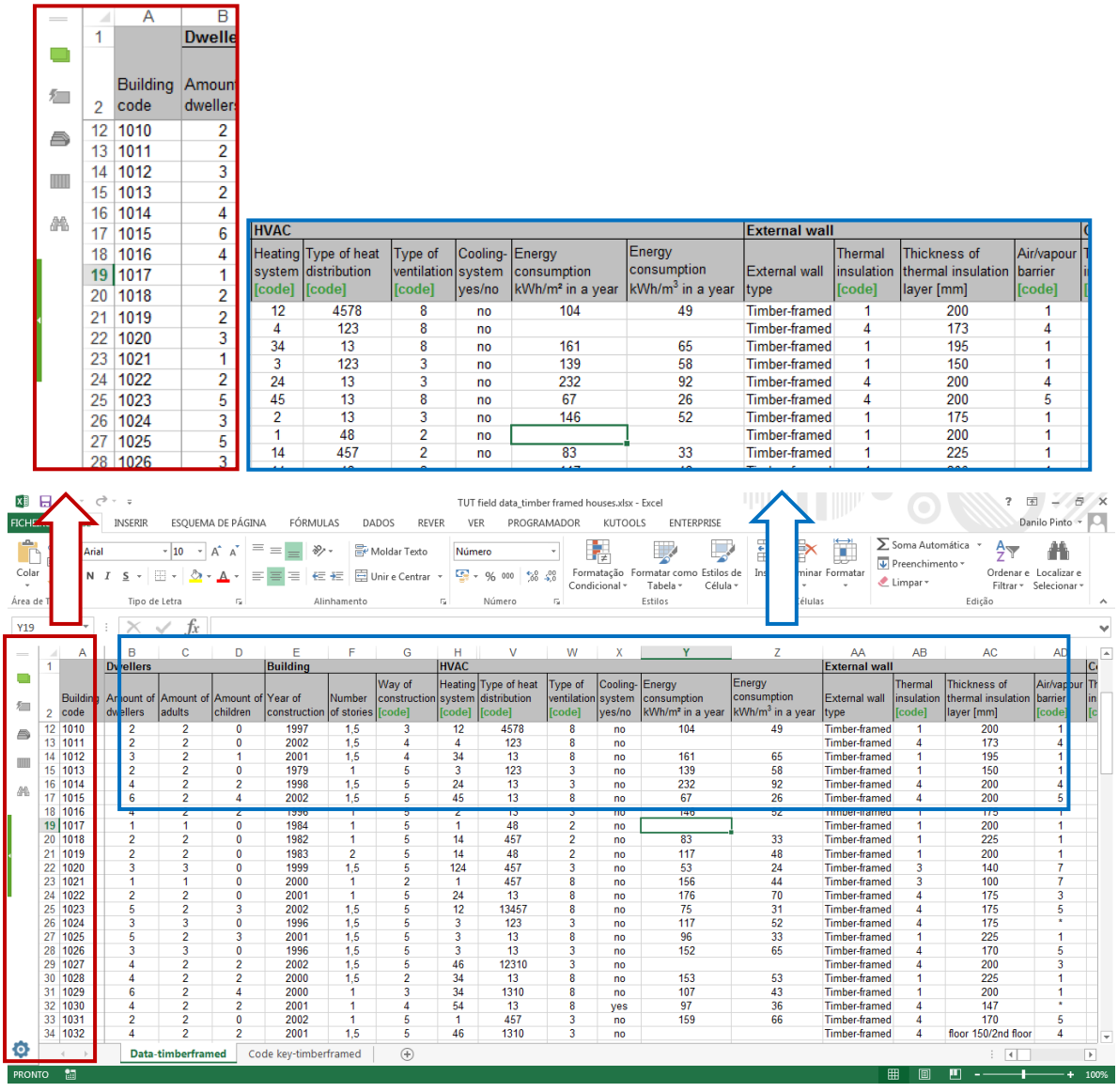


Fig. 3.4. Folha de registo das características das habitações. Em cima à esquerda os códigos representativos de cada habitação. Em cima à direita as diferentes características e respetivos valores ou códigos.

Os dados apresentados na Fig. 3.4 correspondem à totalidade das habitações deste tipo (de 1001 a 1051 e de 2001 a 2053) durante o período correspondente (junho de 2002 a junho de 2004). É de notar que as informações presentes nesta base de dados são relativas a medições anuais por habitação [30].

3.2.2. MORADIAS DE CONSTRUÇÃO PESADA E EDIFÍCIOS DE APARTAMENTOS

Este tipo de habitações, moradias e apartamentos, são constituídas por uma estrutura "pesada" (betão armado e tijolos ou troncos de madeira) e dizem respeito ao segundo grupo referido no início do capítulo. As moradias não têm uma caracterização uniforme entre si, isto é, o tipo de estrutura varia, podendo existir estrutura em madeira tipo tronco, betão armado ou alvenarias resistentes [30].

Fig. 3.5. Modelo de Casas individuais (retirado de [30]).

Fig. 3.6. Edifícios de apartamentos (retirado de [30]).

A temperatura e a humidade relativa interior foram recolhidas do modo apresentado em 3.2.1. Os dados relativos ao exterior foram fornecidos pelo instituto meteorológico da Finlândia, utilizando dados de estações meteorológicas com maior proximidade das habitações [30].

A permeabilidade ao ar e a ventilação foram características medidas de modo semelhante às casas com estrutura de madeira. À semelhança das medições, também as bases de dados criadas são de igual configuração. A organização dos ficheiros é contudo um pouco diferente sendo as habitações divididas por código habitacional e não por ano de medição [30].

3.2.3. ANÁLISE DE PRIMEIRA LINHA

No início deste trabalho, *Data Mining* foi referenciado como uma análise de segunda linha e como tal o conjunto de dados anteriores foi recolhido com um propósito diferente que não DM. No *Annex* 55 são

apresentados alguns projetos ou estudos realizados com estas bases de dados. Este documento apresenta descrições e comparações estatísticas sobre os diferentes parâmetros recolhidos anteriormente [30].

A Ventilação foi descrita e comparada (estatisticamente) com os sistemas de extração de ar existentes para os diferentes tipos de habitação. As medições de temperaturas e humidades apresentam-se também como relevantes para estudo sobre os excessos de humidade e carga estrutural. E por fim foi ainda investigada a permeabilidade do ar em simultâneo com o consumo energético da habitação [30].

O autor Juha Vinha esteve presente ainda em outros artigos que analisam diferentes características das casas finlandesas. Das várias publicações feitas pelo autor, com recurso às diferentes bases de dados apresentadas, neste âmbito importam realçar as seguintes:

- "*Indoor Humidity Loads and Moisture Production in Lightweigth Timber-framed Detached Houses*, tem como objetivo medir e analisar as cargas de humidade interna dos edifícios com estrutura e revestimento de madeira, medindo as temperaturas e humidades interiores e taxas de ventilação. As conclusões do artigo baseam-se em descrever os resultados obtidos relativamente ás variações de temperatura e cargas de humidade internas [31];
- "Airtightness of residential buildings in Finland", cujo objetivo é estudar a estanquidade ao ar das diferentes habitações (moradias e apartamentos) presente atualmente na Finlândia. Este estudo permite concluir que: as casas com teto em estrutura de betão armado têm estanquidade dentro da média ao contrário do que acontece com as que têm teto com estrutura em madeira; as habitações com estrutura e revestimento em madeira dispunham de RPH mais baixo em moradias com múltiplos andares comparativamente às que têm apenas 1. O contrário verifica-se nas de betão; A existência de um espaço de ar entre a laje de solo e o terreno pouco interfere neste aspeto [32];
- "*The effects of ventilation systems and building fabric on the stability of indoor temperature and humidity in Finnish detached houses"*, em que o objetivo consiste no estudo dos efeitos dos sistemas de ventilação e componentes dos edifícios na estabilidade da temperatura e humidade interior. Conclui que a ventilação tem um papel importante na regulação da humidade e temperatura. Este artigo desafia o sistema de classificação da higroscopicidade visto que segundo os seus resultados, ou este sistema de classificação precisa de um melhoramento ou não existem casas absolutamente higroscópicas e não higroscópicas. Por fim conclui ainda que a ventilação tem um papel mais influente no clima interior que as restantes componentes avaliadas [33];
- *"Building leakage, infiltration, and energy performance analyses for Finnish detached houses"*, cujo objetivo foi estudar a relação entre a estanquidade da envolvente de um edifício, a infiltração e o uso energético. O estudo utiliza a mesma base de dados aqui presente para o cálculo de um modelo de uma habitação com estrutura e revestimento em madeira. É através de simulações a esse modelo que se retiram algumas conclusões: a correlação entre a estanquidade da envolvente ao ar com a taxa de infiltração varia de modo linear; A infiltração do ar influencia em 15% a 30% o uso de energia [34].

Como é possível verificar nos parágrafos anteriores, a base de dados já foi utilizada em várias situações diferentes. Estas análises anteriores visam aplicar métodos estatísticos para analisar a correlação entre alguns parâmetros ou certificar algum tipo de método de análise. A partir deste ponto a base de dados deixa de ter a sua primária utilidade, entrando agora o papel de DM.

3.2.4. SELEÇÃO DOS DADOS

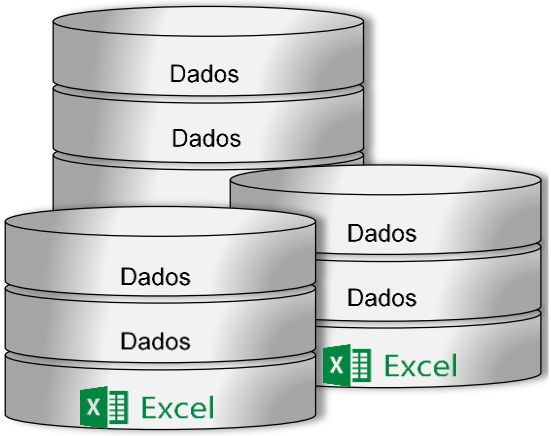


Fig. 3.7. Seleção de dados.

Nos subcapítulos anteriores foram apresentadas as bases de dados disponíveis para a realização deste trabalho. A seleção dos dados a utilizar nessa dissertação tem em conta vários fatores tais como: o objetivo deste trabalho, demonstrar a aplicabilidade de DM no estudo da eficiência energética dos edifícios; a qualidade da base de dados, graças á possibilidade de existir, em demasia, dados corrompidos ou em falta; e o prazo de entrega desta dissertação.

Como apresentado no Capítulo 1, este trabalho tem como fim expor o processo de *Data Mining*, demonstrando que a sua aplicação é possível no campo da Engenharia Civil. A primeira parte do objetivo por si só não tem grande influência na escolha da base de dados, visto que o processo de DM é praticamente constante (os passos são os mesmos, apenas mudando o modo como são executados). A segunda diz respeito a sua aplicabilidade em áreas da Engenharia Civil o que faz dela a mais preponderante na seleção dos dados.

A interseção entre o conforto interior e a eficiência energética tornam-se um fator de peso na seleção dos dados a trabalhar, devido à exigência de paramentos específicos, tais como consumo energético, temperatura interior, entre outros. A qualidade da base de dados é também uma condicionante pois a existência, em excesso, de valores corrompidos ou omissos pode comprometer os resultados das análises. Por último, e com uma grande influência, é o prazo de realização desta dissertação. Este último impõe alguns limites porque no período equivalente à realização deste trabalho torna se bastante difícil uma boa descrição do que é DM e qual o seu potencial e em simultâneo uma análise muito especifica e pormenorizada.

Nas diferentes bases de dados existem dois tipos de informação, aquelas que exigiram registo contínuo e as variáveis de natureza categórica. Os dados armazenados continuamente são referentes às medições horárias de temperatura e humidade relativa. As variáveis nominais são aquelas que caracterizam as diferentes casas.

Geralmente os gastos energéticos das habitações estão ligadas ao aquecimento, à ventilação, à iluminação e aos equipamentos eletrónicos e domésticos. Dos anteriores o aquecimento, a ventilação e a iluminação são aqueles que estão diretamente relacionados com a Engenharia Civil.

O aquecimento, ou conforto térmico por definição presente na ISO 7730, representa o estado psicológico que expressa satisfação com o ambiente térmico, o que é muito difícil de traduzir em parâmetros físicos. O conforto térmico é então estimado tendo em conta o efeito do metabolismo e do vestuário. Para além desses efeitos que se caracterizam de individuo para individuo existem aqueles que são relativos à avaliação do clima anterior. Os parâmetros são a temperatura do ar, a temperatura média radiante, a velocidade do ar e a humidade do ar.

A temperatura radiante corresponde à média das temperaturas das superfícies da envolvente e é o único parâmetro que não dispomos. A temperatura do ar e a humidade são tudo variáveis de que dispomos e que foram consideradas. Existe ainda uma outra informação de que dispomos e que está diretamente relacionada com os três parâmetros anteriores, o isolamento. Este está relacionado com a regulação da temperatura (evitando perdas excessivas) e com o controlo da humidade (prevenindo condensações). É ainda de referir, apenas por curiosidade, que estes parâmetros correspondem ao conjunto de três fatores responsáveis por combater a presença de condensações interiores, isolar, aquecer e ventilar.

A única variável que falta para relacionar a energia com o conforto é o próprio gasto energético. Posto isto as variáveis a considerar são as representadas na Tabela 3.1.

Tabela 3.1 – Variáveis em Estudo

	Temperatura horária anual	Humidade Relativa horária anual	Consumo Energético anual por habitação	Espessura do isolamento	RPH
Unidades	C^{o}	%	KWh/m^2	mm	1/h

Escolhido o conjunto de variáveis a estudar fica apenas a faltar selecionar o conjunto de casas a analisar. Aqui surgem várias opções, a totalidade das habitações, habitações isoladas, apenas as de estrutura de madeira, apenas os apartamentos, entre outras. De um conjunto total de 170 habitações e 56 apartamento resultam um conjunto de 226 habitações, o que levaria uma amostra de dimensão elevada a considerar para este trabalho. Dos vários conjuntos possíveis de se formar a opção recai pelas casas com estrutura e revestimento em madeira (*timber framed houses)*. A razão desta escolha baseia-se no fato de serem idênticas e não variam na sua constituição estrutural ou no revestimento exterior. Deste tipo de habitação apenas serão consideradas as correspondentes ao primeiro ano de medições (2002 a 2003) visto que do segundo ano não existem medições de temperaturas exteriores.

3.3. A ESCOLHA DAS FERRAMENTAS

No Capítulo 2 foram introduzidas algumas ferramentas para a execução do processo de DM. Essas ferramentas dividiam-se em dois tipos, de armazenamento e de análise. No presente subcapítulo é introduzida uma pequena comparação entre elas, bem como a escolha das mais adequadas aos objetivos fixados.

A decisão da escolha de uma dada ferramenta é tomada tendo por base um conjunto de critérios. Os parâmetros de decisão comuns entre os tipos de ferramentas são, a sua disponibilidade (custo), a sua compatibilidade (entre bases de dados e programas de cálculo e vice versa), a facilidade de aprendizagem ou conhecimento prévio e a capacidade de reprodução de resultados (esteticamente). Dos

softwares analisados todos eles, à exceção do Oracle *Database*, são de carater gratuito ou já se apresentam disponíveis através do serviço de aplicações da Universidade do Porto. Simultaneamente ao que acontece com a escolha dos dados a analisar o prazo de realização desta dissertação teve grande influência nas decisões tomadas.

3.3.1 Oracle *Database*, Microsoft Office Excel e Microsoft Office Access

As ferramentas de armazenamento caracterizam-se por dois aspetos fundamentais, a sua capacidade de armazenamento e a facilidade de organizar os dados. A capacidade de armazenamento é um parâmetro que não representa grande importância para este trabalho, visto que a dimensão dos dados em estudo não é astronómica. A capacidade de organização, por outro lado, contem bastante influência na decisão de qual o programa mais adequado ao método de trabalho e tempo disponível.

No início do subcapítulo é referido que Oracle *Database* é o único programa da qual o acesso é condicionado, isto é, apenas está disponível uma versão de demonstração de 15 dias de duração. Este obstáculo foi suficiente para a sua exclusão imediata. Não existia necessidade de adquirir o programa, existindo alternativas gratuitas já disponíveis capazes de armazenar a quantidade de dados em estudo.

A exclusão anterior restringiu a escolha a um dos dois programas do Microsoft Office (Excel e Access). Como referido anteriormente os dois apresentam uma grande diferença, o Excel é uma ferramenta de cálculo que pode ser utilizada como base de dados e o Access é uma ferramenta de armazenamento de dados.

As suas parecenças são bastante evidentes, menus, ícones e até a localização dos mesmos. Seguem todos a mesma linha de *design* adotada pelo Microsoft Office. Contudo existem grandes diferenças a nível de resultados. O Access é uma ferramenta específica de armazenamento e como tal os dados introduzidos ficam automaticamente em formato tabela permitindo a sua organização, rearranjo e filtragem. Os dados introduzidos no Excel ficam num formato mais bruto, isto é, prontos para cálculo e não para gestão. Apesar deste aspeto bruto é possível transforma los em tabelas facilitando a sua organização (Fig. 3.8 e Fig. 3.9).

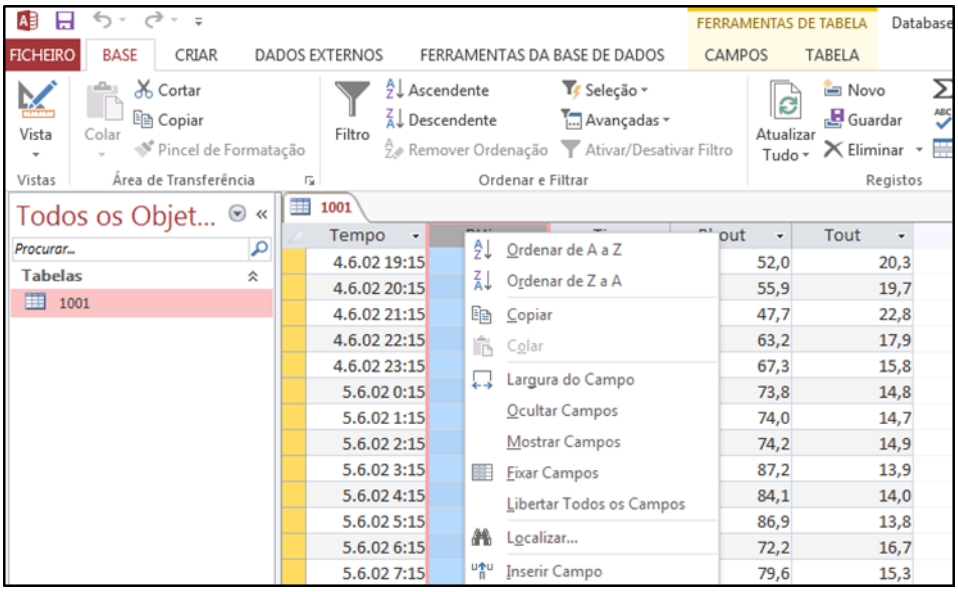


Fig. 3.8. Apresentação dos dados em Access

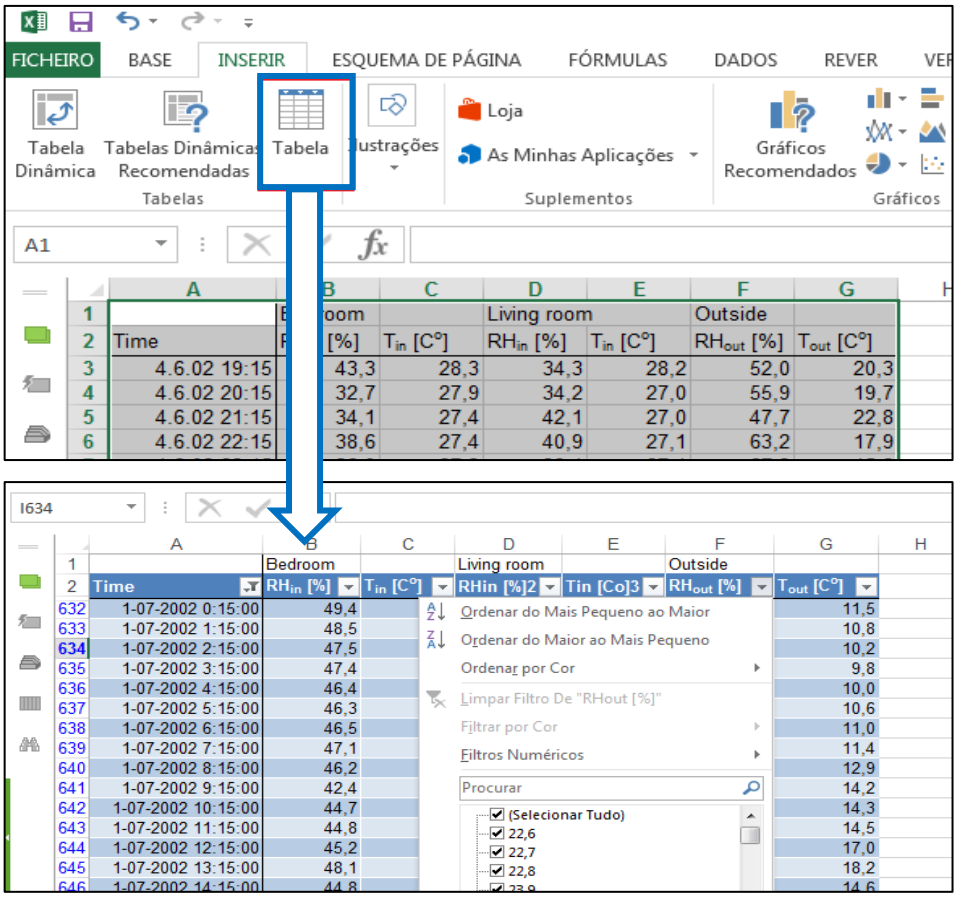


Fig. 3.9. Apresentação dos dados em Excel, em cima no seu estado bruto e em baixo após a execução do comando tabela

Pelas Fig. 3.8 e Fig. 3.9 é possível observar que o Microsoft Office Access demonstra melhor aspeto e maior facilidade na gestão de dados. Como seria de esperar, a sua escolha seria quase inevitável. Contudo os dados iniciais foram todos armazenados em formato Excel (.xlsx) e a sua exportação para Access é demorada. Esta inconveniência fez com que a escolha recaísse no Excel. Aliado a estes factos, o Excel permite ainda a execução de cálculos e a execução rápida de gráficos. Estes dois acontecimentos permitem um ganho de tempo em algumas tarefas analíticas futuras, nomeadamente na fase de processamento de dados. A Tabela 3.2 apresenta um resumo das características avaliadas na escolha da ferramenta.

Tabela 3.2 – Tabela Síntese

	Microsoft Office Excel	Microsoft Office Access	Oracle *Database*
Disponibilidade	Gratuita	Gratuita	Não gratuita
Compatibilidade (com os programas de análise)	Sem problemas	Sem problemas	-
Intuitivo	Sim	Sim	-
Experiencia prévia do utilizador	Sim	Não	Não
Apresentação	Média	Boa	-
Formato original dos dados	Sim	Não	Não
Ferramenta escolhida	Microsoft Office Excel		

3.3.2. FERRAMENTAS DE ANÁLISE

À semelhança do que acontece nas ferramentas de armazenamento também as de análise passaram por um processo de comparação entre si. Como referido anteriormente existem dois grupos distintos, os programas estatísticos (e matemáticos) e as "caixas negras". O primeiro grupo é constituído pelas ferramentas MATLAB, RStudio, STATISICA e IBM SPSS, o segundo é constituído por Weka e RapidMiner.

Esta comparação divide-se em duas etapas. Na fase inicial, abrange apenas os programas estatísticos e matemáticos. Os programas tipo "caixa negra" serão introduzidos numa etapa mais avançada. Esta separação ocorre porque estes programas são quase desconhecidos na área da Engenharia Civil.

3.3.2.1 MATLAB, RSTUDIO, SPSS E STATISTICA

Os programas MATLAB, RStudio, SPSS e STATISTICA apresentam uma diferença capaz de os dividir em dois grupos. Essa diferença diz respeito ao seu modo de trabalhar. Os *softwares* SPSS e STATISTICA funcionam com uma interface gráfica do utilizador (GUI), enquanto que RStudio apresentam uma interface do tipo linha de comandos (CLI). MATLAB apresenta os dois tipos de interface mas nesta análise será considerado como um programa de CLI, porque é neste formato que as suas capacidades são maximizadas. A diferença entre ambos está presente no tipo de facilidade com que o utilizador trabalha. Para uma melhor compreensão basta observar o próprio sistema operativo Windows. Na fase inicial da computação todas as ações elaboradas nos computadores eram através de comandos por extenso (CLI). À medida que a tecnologia evolui foi possível atribuir ícones a esses comandos, nascendo assim GUI.

Através da Fig. 3.10 é possível perceber a grande diferença entre estas duas interfaces. A base de dados utilizada para comparação diz respeito às rolhas de cortiça e foi disponível por [7].

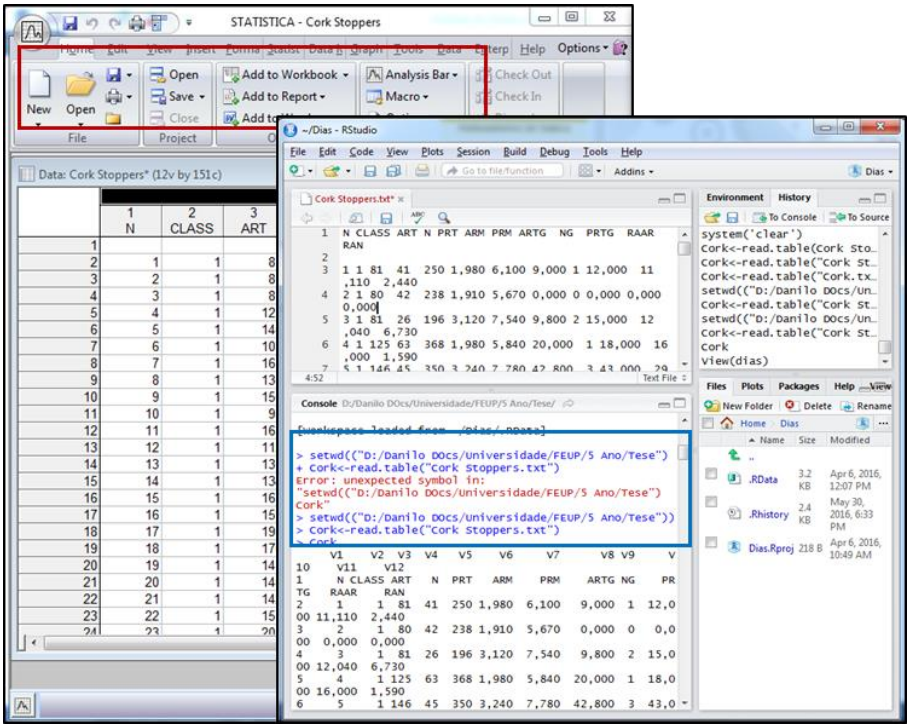


Fig. 3.10. A vermelho o modo de introdução de dados no STATISTICA (GUI) e a azul o modo de introdução de dados no RStudio (CLI).

Considerando ainda a Fig. 3.10 pode-se retirar muitas outras conclusões entre a diferença dos dois tipos de programas. O modo de apresentação dos dados em análise é bastante diferente. No SPSS e STATISTIC os dados a trabalhar são apresentados em células tal como nas bases de dados origem, contrariamente ao que acontece com MATLAB e RStudio que apresentam os dados de forma matricial. Apesar das diferenças presentes até agora, ainda não é possível avançar com a escolha de um favorito. Para tal é necessário alargar a comparação a níveis mais específicos.

A escolha da ferramenta indicada tem que garantir um ganho de tempo na sua utilização e uma aplicabilidade quase completa da metodologia adotada. Com o método de trabalho definido é possível prever quase todas as funções a serem utilizadas. Sabendo os principais métodos adotados em cada passo é possível comparar as ferramentas. A Melhor forma de o fazer é através de uma pequena análise multicritério. Este tipo de análise não foi aplicado às ferramentas de armazenamento porque a comparação era apenas entre duas delas e a quantidade de parâmetros a considerar menor.

A Tabela 3.3 descreve os resultados de uma pequena análise multicritério elaborada. Esta análise é divida nos quatro parâmetros visíveis na primeira coluna, contendo cada um deles um peso característico para ponderação.

Tabela 3.3 – Análise multicritério

	MATLAB	RStudio	SPSS	STATISTICA
Linguagem (5%)	2	1	3	3
Reprodução da Metodologia (40%)	3	3	2	3
Apresentação de resultados (25%)	2	1	3	2
Experiencia prévia do utilizador (35%)	1	0	2	0
Pontuação	2,15	1,5	2,4	1,85

Observa-se que a "Linguagem" apresenta um peso baixo nesta análise. Isto acontece porque está indiretamente relacionada com a "Experiencia prévia do utilizador". Este relacionamento acontece porque ao considerar que o utilizador tem experiencia sobre a ferramenta que usa o tipo de linguagem torna-se um pouco irrelevante, contudo não desprezável. Por exemplo, se dois utilizadores possuírem o mesmo grau de conhecimento de STATISTICA e RStudio é provável que aquele que dispõe de uma interface gráfica de utilizador consiga maior rapidez na execução de comandos (Fig. 3.10). Somando os dois critérios obtemos 40% da ponderação total, o que representa bastante. Esta percentagem foi assim atribuída graças ao fator tempo (5%+35%) pelas razões anteriormente enunciadas na escolha da base de dados. Passa então a ser considerado um "ganho de tempo" o facto de o utilizador conhecer a ferramenta.

Faltam apenas explicar dois parâmetros, a "Reprodução da Metodologia" e a "Apresentação de Resultados". A reprodução da metodologia corresponde a 40% desta ponderação. Tendo em conta o objetivo desta dissertação é de esperar que a reprodução da metodologia seja tão importante quanto o fator tempo. O parâmetro "Reprodução da metodologia" tem como base o fato cada programa, em exclusivo, ser capaz de cumprir todos os passos da metodologia de um modo eficaz e eficiente. O critério "Apresentação de Resultados" representa a apresentação, aspeto e clareza do resultado de um dado comando de cada *software*. Este contem um peso médio porque apesar de DM ser um processo muito gráfico (tabelas, gráficos, diagramas, resultados,…), todo o conjunto de elementos deste género é dotado de uma explicação, pelo menos ao longo deste trabalho.

As pontuações na Tabela 3.3 distribuem-se entre um máximo de 3, fácil, e um mínimo de 1, difícil, existindo ainda o atributo 0, desconhecimento ou exclusão total. As classificações foram atribuídas como resultado de pequenos testes efetuados. Os testes efetuados utilizam a base de dados referida no início do subcapítulo e a demonstração de resultados seguinte tem suporte no livro respetivo [7].

A Pontuação sobre o primeiro e o último parâmetro da análise multicritério não necessitam de grande explicação à margem daquela fornecida na sua respetiva apresentação. Enquanto em GUI a introdução dos dados é muito intuitiva e fácil, em CLI é necessário escrever um comando, mais demorado e mais suscetível a erros de introdução (Fig. 3.10). A distribuição de classificações sobre a "Reprodução da metodologia" é deste modo devido a duas diferenças principais. A primeira diz respeito ao facto de que o SPSS é mais restrito na utilização de certos métodos contrariamente ao MATLAB e RStudio. Os *softwares* que utilizam CLI dispõem de mais liberdade porque os comandos são introduzidos pelo utilizador (por vezes com alguns ajustes) e não já intrínsecos ao programa. Esta capacidade traduz-se, não apenas para a metodologia adotada, mas para todos os métodos presentes em DM (alguns

apresentados no Capítulo 2). Esta desvantagem, que as ferramentas que dispõe de GUI sofrem, é colmatada pelo programa STATISTICA (ao nível do processo de DM como visível na Fig. 3.10). É então esta a segunda diferença, entre o SPSS e o STATISTICA, que resulta nas pontuações indicadas na Tabela 3.3.

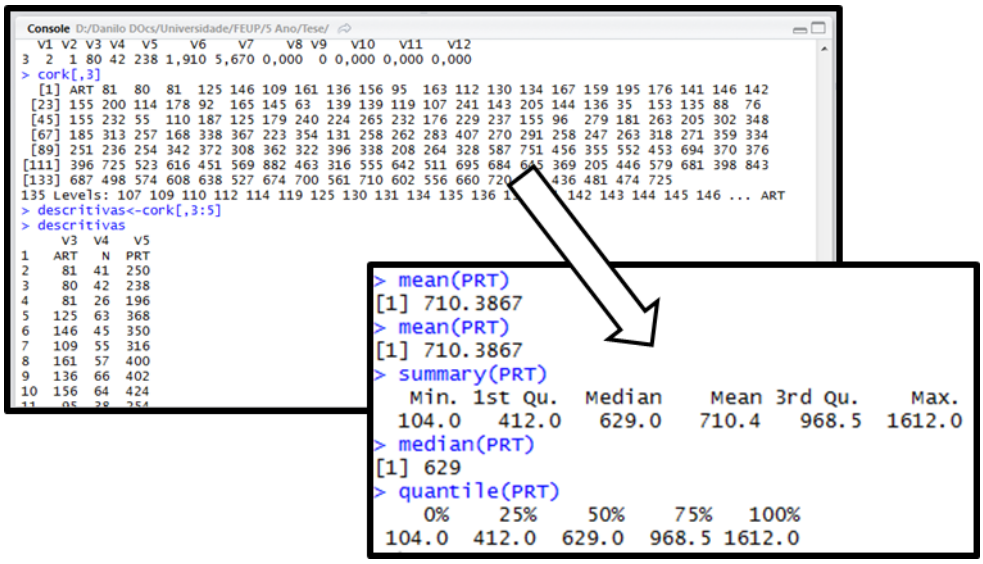


Fig. 3.11. Estatísticas descritivas em RStudio.

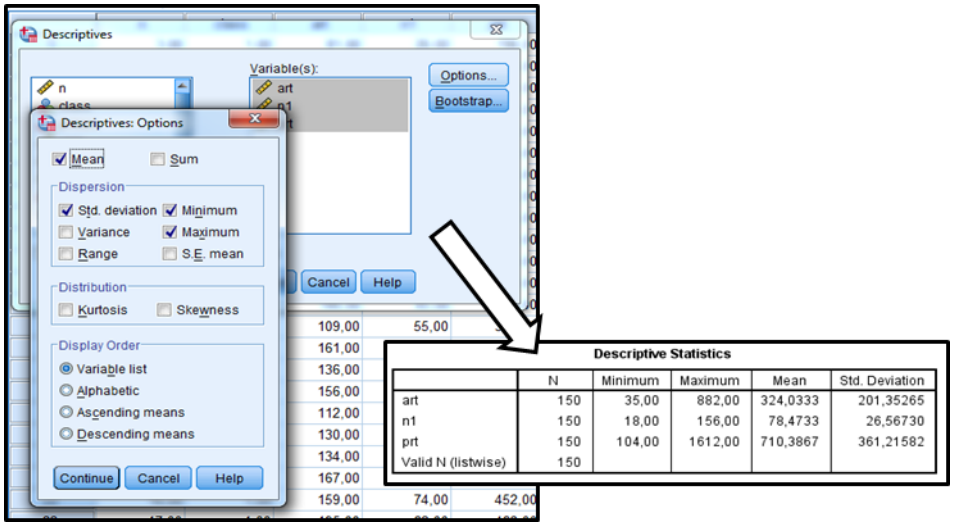


Fig. 3.12. Estatísticas descritivas em SPSS.

A classificação do critério "Apresentação de Resultados" é determinada através do grau de grafismo associado e da sua capacidade de manipulação. A pontuação presente na tabela 3 tem algumas particularidades interessantes. RStudio (Fig. 3.11) é considerado o pior programa em termos de apresentação quando o MATLAB funciona do mesmo modo. MATLAB, por sua vez, utilizando CLI consegue igualar – se ao programa STATISTICA. Por fim SPSS (Fig. 3.12) tem-se como o melhor programa em termos de apresentação.

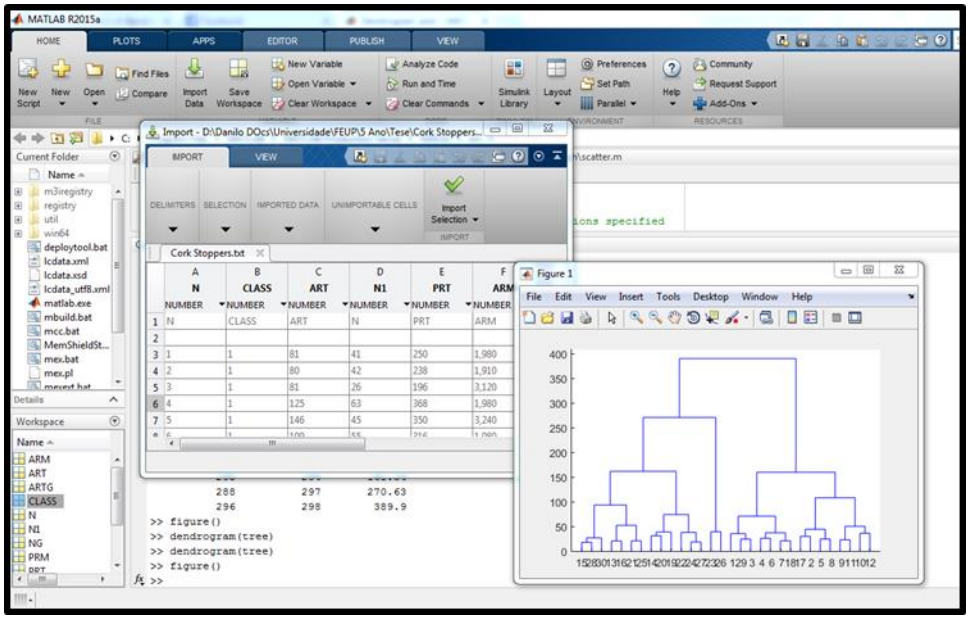


Fig. 3.13. Dendrograma (ART) em MATLAB

No canto inferior direito da Fig. 3.13 é possível observar um dendrograma, o que representa outra desvantagem relativamente aos *softwares* CLI. Neste tipo de programas a formatação dos gráficos torna-se mais difícil e por vezes os valores apresentados encontram-se em elevado número obrigando a utilização de escalas menos apropriadas. Este fenómeno apenas se verifica em MATLAB e RStudio.

Avaliando os resultados da análise multicritério em função do estudo anterior é possível retirar algumas conclusões. O SPSS é o favorito para esta metodologia, mas apenas devido ao "ganho de tempo" anteriormente falado. É também conclusivo que os outros programas apresentam um maior potencial em análises de DM mais complexas. RStudio apresenta um grande potencial para metodologias detalhadas e demoradas, bem como MATLAB e STATISTICA, contudo Rstudio é gratuito e apresenta um nível de desempenho não muito diferente.

3.3.2.2 SPSS, RAPIDMINER E WEKA.

Esta secção corresponde à segunda fase de seleção de uma ferramenta de análise estatistica. SPSS foi a ferramenta selecionada entre os programas matemáticos e estatísticos. Falta agora uma pequena comparação entre as vantagens da sua utilização perante os programas tipo "caixa negra" ou o contrário.

No Capítulo 2 foi apresentada uma pequena descrição sobre os programas tipo "caixa-negra". A grande diferença entre este tipo de programas e os anteriores (SPSS, MATLAB, RStudio e STATISITCA) é que a execução das várias ferramentas é direta e não depende de um conjunto de procedimentos. Em SPSS a utilização de métodos, como por exemplo clusterização, depende de um conjunto de procedimentos e de parâmetros definidos pelo utilziador. Neste tipo de *software* o modo de trabalho é diferente, visto que os parâmetros já lhe são intrínsecos permitindo apenas a aplicação do método. Tendo em conta o objetivo desta dissertação existem dois fatores para a não utilização deste tipo de ferramentas. O primeiro é relativo ao fato de que este tipo de programas dispõe da sua própria linguagem e modos de funcionamento, sendo necessário um estudo mais demorado sobre os mesmos. O segundo fator é caracterizado pela necessidade de recorrer a uma validação dos resultados obtidos. Esta acreditação

passaria pela comparação com outro programa que já tenha demonstrado resultados na aplicabilidade de DM. Apesar da sua exclusão apresentam-se as Fig. 3.14 e Fig. 3.15.

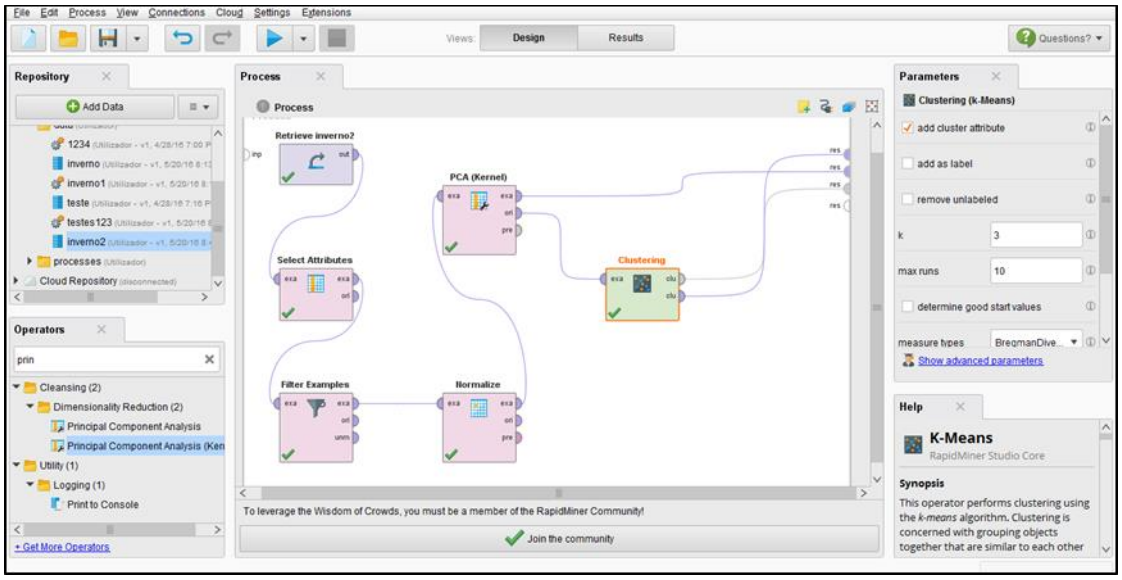


Fig. 3.14. Interface de Rapidminer.

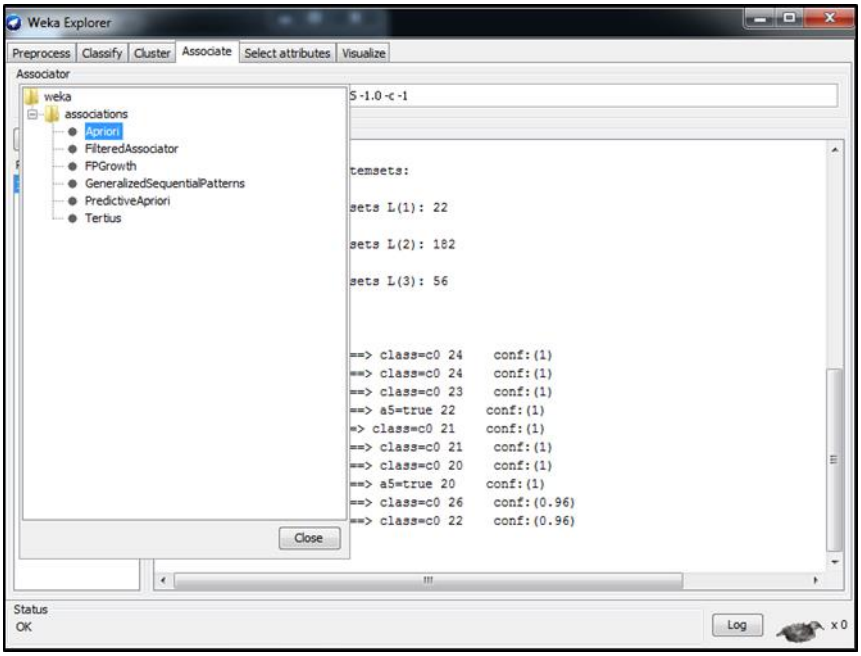


Fig. 3.15. Interface de Weka.

3.4. SPSS E EXCEL

Uma vez apresentada a metodologia e escolhidas as ferramentas resta demonstrar a sua aplicabilidade. Como referido anteriormente, SPSS e Excel, representam uma dupla capaz de reproduzir com eficiência esta metodologia. Nesta secção são demonstrados os diferentes procedimentos para a obtenção dos

resultados presentes no próximo capítulo. Apresentam-se três das cinco etapas que este procedimento é constituído. Note-se que a primeira fase, seleção de dados, já foi demonstrada em 3.2.4 e que a ultima, interpretação de resultados, está presente nos capítulos seguintes. Existindo dois tipos de variáveis disponíveis (qualitativas continuas e qualitativas nominais) são necessárias abordagens diferentes no contexto de DM.

Os dados inicialmente encontram-se em estado bruto, e caracterizam-se por difícil leitura, sem um período uniforme de recolha de dados e com algum ruido. O primeiro passo consiste em formatar a apresentação da base de dados. A Fig. 3.9 também apresenta a possibilidade de formatação de dados em tabela para uma maior facilidade na sua organização. Com o formato definido, diminuindo assim os erros de leitura entre *softwares*, o passo seguinte consiste na uniformização do período de estudo.

Dispondo em Excel as datas correspondentes às medições de cada habitação é possível determinar o número de dias que cada uma esteve sob recolha. Posteriormente à sua identificação é necessário encontrar um intervalo que contenha boas condições para analisar o problema em estudo. É de notar que sendo o objetivo estudar a relação da eficiência energética com o conforto, os meses mais cruciais para esta análise são os de inverno e verão.

Por fim, encontrado o período de análise, resta colmatar os dados em falta. Uma vez já formatados os dados podem ser introduzidos em SPSS e por sua vez encontrar falhas na informação. O SPSS tem uma função específica capaz de encontrar e substituir essa informação. Este processo permite ainda estimar os valores em falta, através do método de "expectation maximization".

Com o fim do processamento de dados dá-se início à transformação dos mesmos. Esta 3ª etapa prepara a base de dados para a formação de padrões. Tendo em conta a configuração inicial dos dados, medições horárias durante um ano, a carga de informação é em demasia para poder ser analisada. Como tal é necessário requerer a agregação da mesma. O cálculo de algumas medidas descritivas, tais como a média e o desvio padrão dos parâmetros horários, para cada habitação ao longo dos meses em estudo é um modo de tentar concentrar mais a informação. Mas esse processo de agregação pode ser ainda mais completo, através de um estudo sobre as correlações entre as diferentes variáveis. Em concordância com o resultado das correlações pode-se perceber até que ponto é possível reduzir o número de variáveis. É também adequado aplicar a análise das componentes principais de modo a reduzir o tamanho da amostra em estudo, sem comprometer a qualidade da informação.

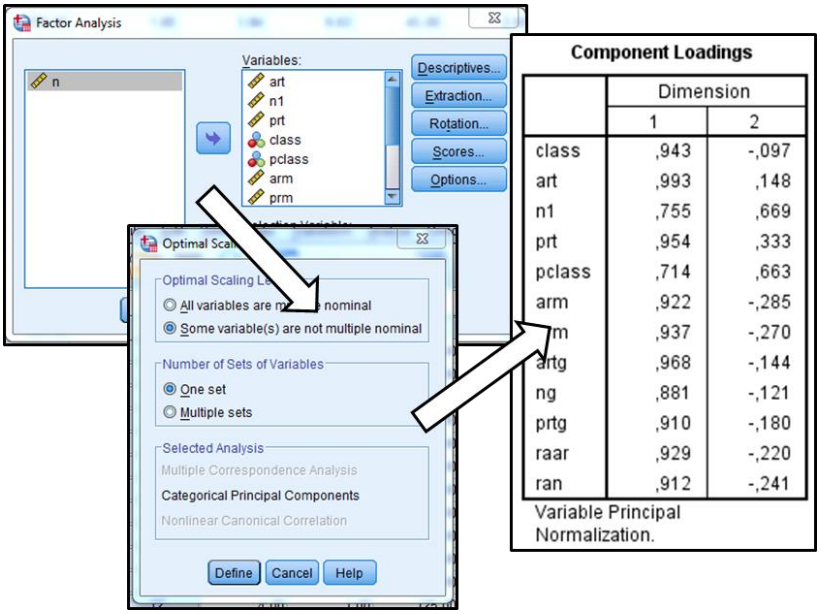


Fig. 3.16. Análise dos componentes principais em SPSS.

A Fig. 3.16 apresenta os passos a seguir para aplicação de análise de componentes principais. Primeiramente o utilizador deve optar por elaborar uma análise de fatores para perceber a quantidade de informação explicada por cada componente. Conhecendo o número ideal de componentes, é só aplicar o ACP. Dos vários métodos de normalização, o selecionado neste trabalhado é o método da variável principal. O resultado do ACP está demonstrado na tabela Fig. 3.16, em que foram criadas duas novas variáveis (componente 1 e 2) que descrevem quase a totalidade da informação presente nas variáveis iniciais. Este método de redução irá permitir que o número de variáveis seja reduzido drasticamente perdendo pouquíssima informação.

A última etapa aqui presente é a formação de padrões e classificação da informação de que dispomos. Com o número de variáveis já reduzido é mais fácil encontrar informações escondidas nos dados. Esta fase começa com a formação dos *clusters* relativamente às componentes encontradas anteriormente. Como referido anteriormente um dos problemas da análise de *clusters* é a determinação do número de grupos a formar. Para tal, o caminho escolhido nesta dissertação passa por aplicar um método hierárquico para a formação de *clusters* de modo a determinar o número ideal de grupos e por sua vez aplicar o método "*K-means*". Deste modo é possível perceber com um maior detalhe os grupos formados e a sua composição. Nesta fase entra um pequeno tratamento qualitativas nominais. As que servirão para posterior classificação. Estas variáveis são relativamente fáceis de analisar, pois estudando as frequências relativas acumuladas é possível determinar qual a melhor classificação a atribuir. Cruzando os grupos formados com a classificação escolhida, constrói-se a árvore de decisão e analisa-se.

4
ANÁLISE DE RESULTADOS E DISCUSSÃO

4.1. PROCESSAMENTO DE DADOS

Os resultados relativos à aplicação da segunda etapa da metodologia adotada apresentam-se em seguida. Após a formatação, já falada anteriormente, os dados ficaram mais fáceis de trabalhar mas continuou a ser necessária uma análise sobre a qualidade da informação presente. A remoção de *outliers* é um processo que está presente em várias etapas de DM. A primeira a que este é aplicado é na uniformização do período de tempo em análise. A base de dados dizia respeito a medições horárias durante um ano de um conjunto de casas. Então deu-se início a um estudo sobre a contagem de dias em que existiram registos. Essa contagem de dias permitiu retirar dois tipos de resultados, quais os *outliers* e qual o melhor período para análise, mais completo.

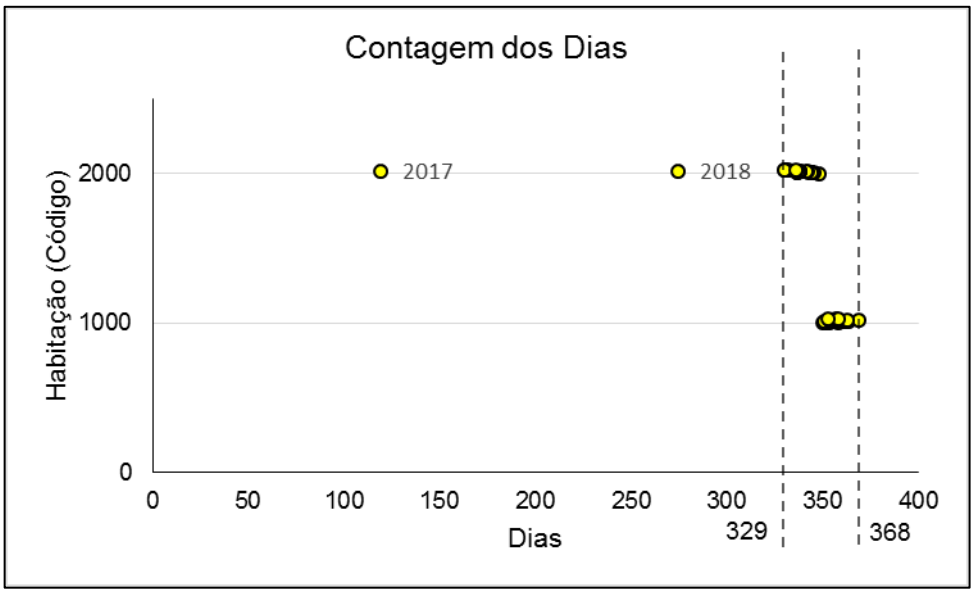


Fig. 4.1. Representação do nº total de dias com recolha horaria de temperatura.

O gráfico de dispersão anterior (Fig. 4.1) demonstra a existência de dois *outliers*, habitações 2017 e 2018, que serão excluídas do estudo. Tendo em conta a descrição da base de dados, o número de dias deveria rondar os 365, o que não se verifica nestes dois casos. Verifica-se também que o período de

medições enunciado não corresponde à realidade. Entre as duas linhas a tracejado é possível verificar que o número de dias varia entre 329 e 368. É na mancha presente dentro deste intervalo que será necessário escolher um período de tempo uniforme. Apesar de praticamente para todas as habitações, se iniciarem no sexto mês as medições, o dia correspondente à primeira variou bastante. Desse modo a melhor opção para iniciar a análise foi considerado ser o primeiro dia de Julho. Determinado o início fica a faltar o fim do período de estudo. Este apresenta uma maior heterogeneidade (término das medições em dias diferentes dos meses de maio e junho) e por isso a melhor opção foi restringir o período de estudo até ao último dia de abril. O intervalo de tempo considerado, entre 1-07-2002 e 30-04-2003, abrange os meses normalmente mais críticos para este tipo de análise (meses de verão e de inverno).

Com o número de casas a analisar já delimitado continuou-se o processamento dos dados. Foi demonstrado na Secção 3.2.1 o aspeto e organização da base de dados. Esta apresentava medições horárias ao longo de cada dia durante alguns meses. É de fácil compreensão que tal dimensão tornava a análise demasiado pesada e complexa. Esse problema foi resolvido através do cálculo das médias e desvios padrões horários da temperatura e humidade relativa de cada habitação. Assim, é possível descrever a informação presente nos dados recolhidos de uma forma mais concisa mantendo a qualidade necessária.

Com o período de análise correspondente a 10 meses foi feita uma análise estatística descritiva para perceber a homogeneidade da informação. Após alguma análise e discussão de resultados verificou-se que por falta de consistência entre resultados ficava difícil um estudo tão abrangente, optando-se pela divisão em estações do ano. Considerando o intervalo de tempo selecionado é percetível que a primavera não se encontra completamente descrita. Este problema aliado ao fato de ser uma estação "intermédia", assim como o outono, não apresenta parâmetros meteorológicos homogéneos ao longo da sua duração, levou à sua exclusão na análise. Para além dos fatos anteriores citados, foi ainda tomado em conta que, em relação à análise de conforto interior, os meses de inverno (principalmente) e verão são aqueles que podem causar um maior interesse e impacto. Este conjunto de considerações levou a que apenas fossem tomadas em conta no estudo as medições relativas aos meses de inverno e verão. Segundo o clima frio da Finlândia a distinção entre inverno e verão fica complicada contudo é adotado a distinção convencional dezembro a fevereiro corresponde ao inverno e apenas o mês de agosto para o verão.

Tendo em conta a leitura do parágrafo anterior podemos concluir que a homogeneidade da amostra corresponde a uma palavra-chave nesta fase. É importante que a amostra seja coerente para que os resultados possam ser fidedignos mas também é importante que a homogeneidade não seja excessiva, porque pode afetar a análise de *clusters*. Torna-se então necessário encontrar um ponto de equilíbrio, a amostra não pode ter valores demasiado dispersos nem demasiado homogéneos. Este ponto de equilíbrio apenas pode ser encontrado através de várias iterações relativas à metodologia apresentada. Esta análise de homogeneidade das amostras pode ser vista através de um gráfico de dispersão (Fig. 4.2). Os meses relativos ao inverno são dezembro, janeiro e fevereiro enquanto os que nesta análise dizem respeito ao verão são os de agosto e julho. No entanto julho foi excluído dos meses representativos do verão porque é correspondente a um mês de férias na Finlândia. Com este fato percebe-se que a sua introdução nesta análise pode comprometer alguns resultados visto que as rotinas dos ocupantes são diferentes da normalidade.

À medida que esta descrição ia continuando foi detetado uma condicionante que levou à exclusão das medições das temperaturas e humidades relativas exteriores da nossa amostra. Este acontecimento surgiu uma vez que as variáveis não aparentavam grande variabilidade. Este fato levou à consideração de que as diferentes habitações estavam distribuídas por um raio geográfico relativamente curto, uma

vez que não se verificaram diferenças significativas de temperaturas e humidades relativas exteriores entre habitações.

Em estatística, a média amostral representa um valor que descreve a generalidade dos casos de estudo. Contudo o seu cálculo pode não ser suficiente para caracterizar devidamente a informação. Esta situação surge porque a média foca-se apenas em representar o valor central de um certo conjunto de dados ficando suscetível a sofrer influência de valores que estejam em extremos do conjunto de dados. Para que seja então possível caracterizar devidamente estes dados torna-se necessário a introdução de uma medida de dispersão, o desvio padrão. O desvio padrão é uma medida responsável por avaliar a dispersão presente na amostra, isto é, o afastamento dos valores máximos e mínimos perante os valores centrais ou média. Anteriormente foi referida a importância da coerência neste tipo de análise para que os resultados possam ser os mais autênticos possíveis. Essa coerência pode ser avaliada através do estudo da dispersão presente na amostra. Tendo as descrições feitas até este ponto, é de fácil compreensão que quanto mais baixo os valores dos desvios padrões menor será a variabilidade presente nos dados e por consequência maior será a sua homogeneidade. Seguindo um padrão semelhante de homogeneidade ao longo dos meses, pode-se assumir que os dados são coerentes garantindo assim uma melhor eficácia de resultados.

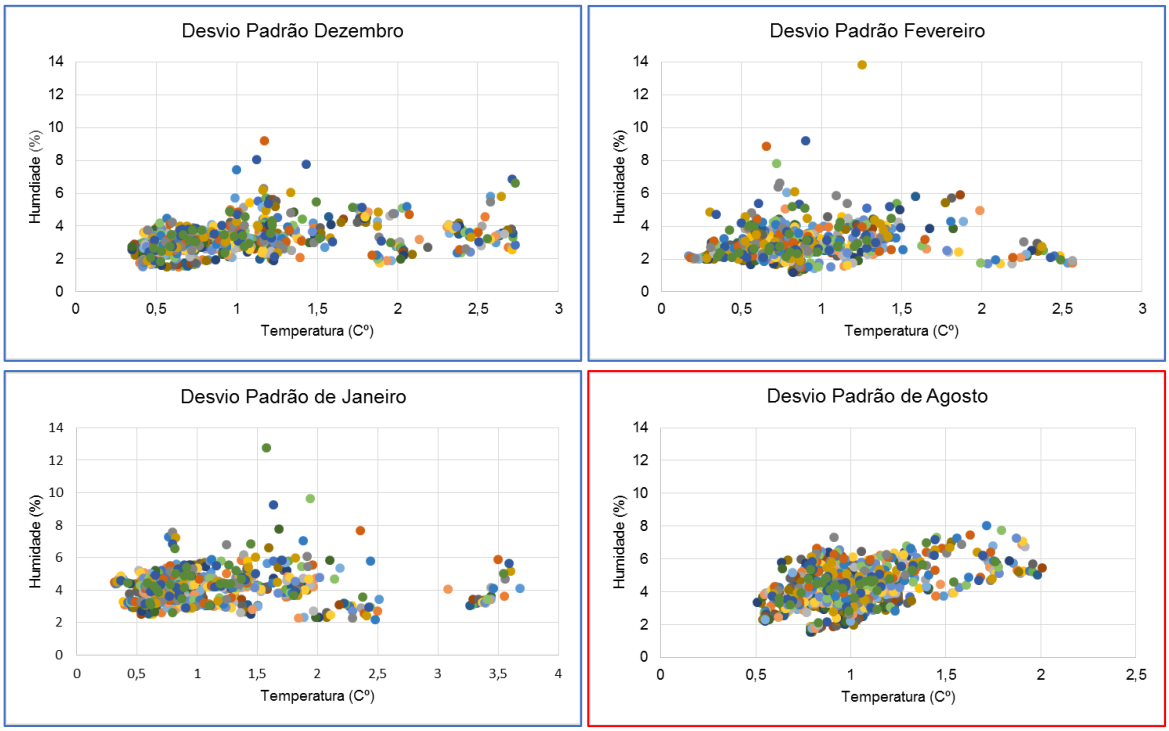


Fig. 4.2. Gráficos de Dispersão de Desvio Padrão nos meses de inverno (azul) e verão (vermelho).

A Fig. 4.2 apresenta os gráficos de dispersão cruzando os dados relativos aos desvios padrão horário (da humidade relativa e da temperatura) nos meses de inverno e verão. Os meses de inverno apresentam uma coerência bastante alta, onde é possível visualizar a proximidade entre os contornos das manchas. Para além deste fenómeno é possível ver que a maior parte da informação apresenta um desvio padrão relativamente baixo, tanto a nível de humidades como temperaturas. Ora este fato leva a que a homogeneidade destes conjuntos seja grande, o que pode levantar dificuldades na análise de *clusters*

caso seja demasiado alta, contudo se o contrário se verificar pode comprometer a fiabilidade dos padrões. Ao observar o gráfico relativo ao mês de verão podemos verificar o contrário. O conjunto de dados selecionado apresenta uma homogeneidade mais alta ao contrário do que se verifica durante os meses de inverno.

Com esta análise foi então possível concluir que os dados relativos aos meses de inverno apresentam uma grande compatibilidade entre si e que os meses de verão apresentam uma homogeneidade mais acentuada. Este fato leva a que a análise dos meses de inverno seja mais acessível e com maior probabilidade de melhores resultados.

Com o período de análise finalmente estabilizado procedeu-se a um estudo sobre a amostra em questão. Esse estudo teve como objetivo fazer uma caracterização dos dados, isto é, perceber o seu comportamento. Esta caracterização aconteceu para que fosse possível um maior envolvimento e entendimento sobre as informações que estes dados transmitem. Este processo contou com o estudo da variação da temperatura e humidade relativa ao longo dos meses em estudo. Através de algumas medidas estatísticas descritivas e algumas representações gráficas (*Box-Plot)* é possível observar mais concretamente a distribuição da amostra em estudo.

Como referido até este ponto, a análise desta base de dados tem como foco principal o estudo da temperatura e humidade relativa interior. Estas duas variáveis são boas representantes do nível de ambiente interior existente nas diferentes habitações. Em seguida apresentam se as variações destes dois parâmetros.

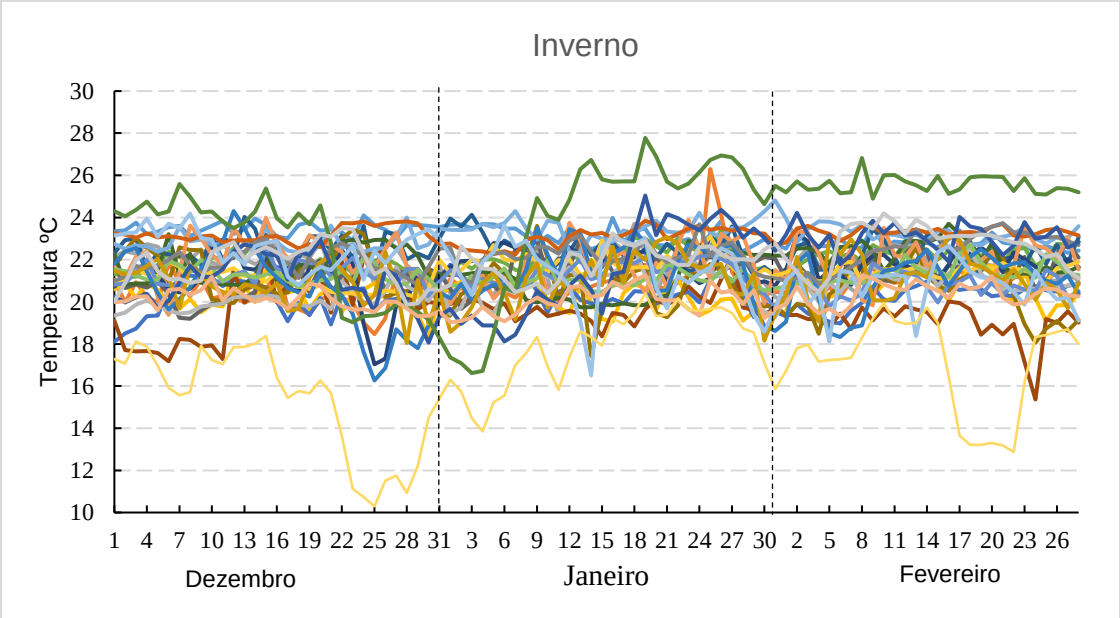


Fig. 4.3. Variação da temperatura média diária ao longo do inverno.

Na Fig. 4.3 é apresentada a variação da temperatura relativa aos meses de dezembro, janeiro e fevereiro. Cada linha, de coloração diferente, representa uma casa e o seu percurso corresponde às temperaturas médias encontradas ao longo de cada dia do mês. Observando o gráfico o primeiro aspeto a realçar é a existência de duas habitações (representadas a amarelo e verde) cujo traçado não corresponde à normalidade da amostra. As variações apresentadas por ambas as habitações são de elevado grau

podendo isto indicar que se caracterizam como *outliers* desta amostra. Analisando a restante amostra é possível observar que a variação nas diferentes habitações segue o mesmo perfil. Observando com atenção é verifica-se que as temperaturas seguem uma distribuição quase cíclica, isto é, vai apresentado picos máximos e mínimos ao fim de um número de dias aproximadamente constante. Após o cruzamento com o calendário foi possível perceber que esses ciclos correspondem a períodos que diferem entre fim-de-semana e semana laboral. No geral, as habitações apresentam uma subida de temperatura ao fim de semana e uma ligeira descida durante a semana. Esta situação pode ser explicada pelo maior tempo de permanência em casa por parte dos habitantes ao fim de semana levando a um aumento natural da temperatura. É ainda curioso observar que a temperatura sofre uma descida mais acentuada às quartas-feiras e é anómala durante os feriados natalícios, talvez por causa da tradição de passar o Natal em família. A Fig. 4.4 mostra a variação da média da temperatura diária ao longo do verão relativo ao mês de agosto.

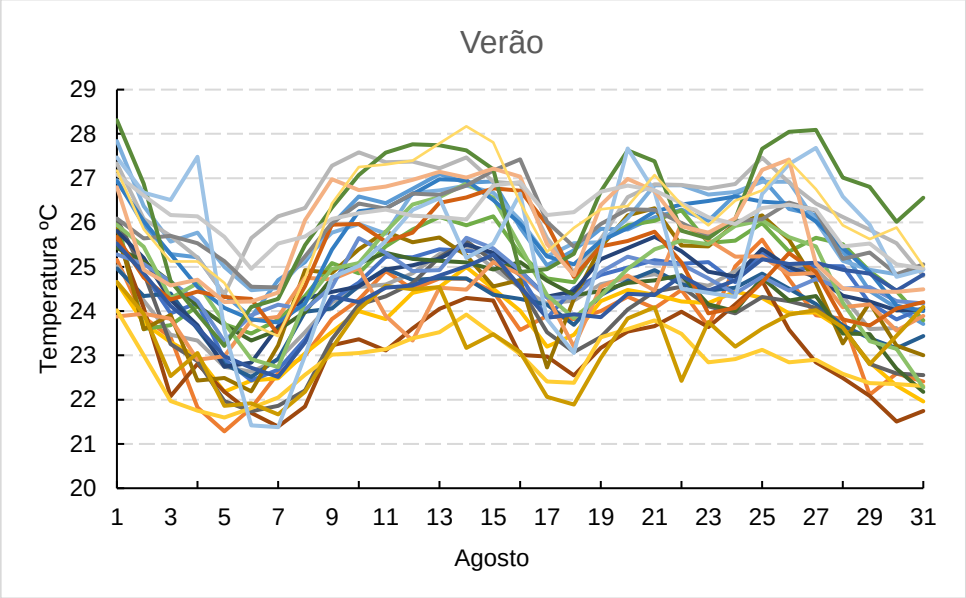


Fig. 4.4. Variação da temperatura média diária ao longo do verão.

Observam-se algumas semelhanças com os meses de Inverno. Essas semelhanças estão presentes na distribuição das temperaturas em função do tempo, seguindo o mesmo formato de altos e baixos cíclicos, e nas extremidades que continuam a ser representadas novamente pelas curvas a verde e amarelo. Apesar de a consideração de *outlier* não ser tão evidente, estas habitações pelo fato de estarem muito próximas do extremo da amostra é um ponto a favor dessa consideração. Apesar das semelhanças este gráfico demonstra resultados contrários ao anterior, isto é, a descida de temperatura ocorre ao fim de semana e não durante a semana. A Fig. 4.5 apresenta a variação da temperatura exterior ao longo dos meses de inverno para efeitos de comparação da distribuição com as temperaturas interiores.

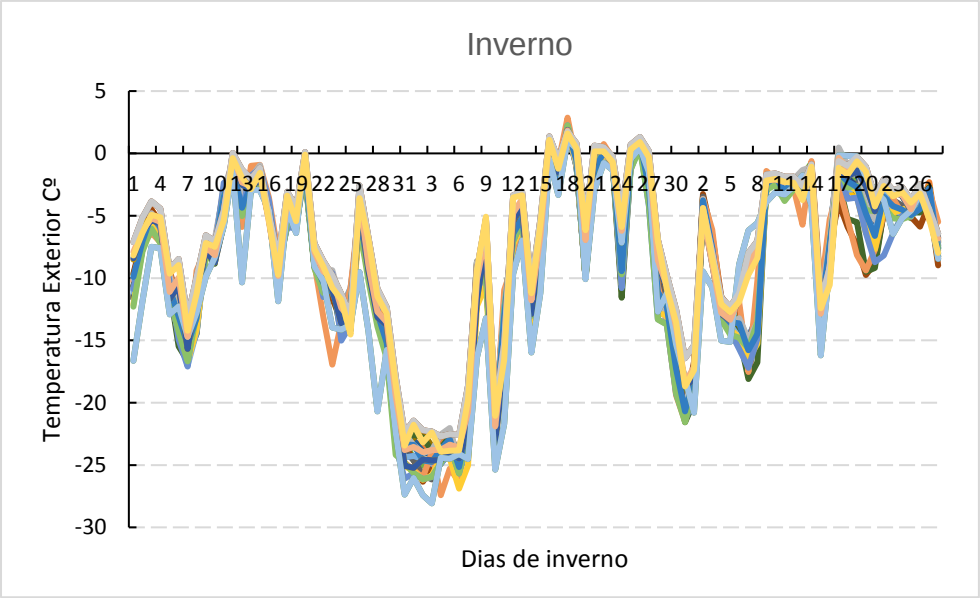


Fig. 4.5. Variação da temperatura média diária exterior ao longo do inverno.

É visível uma distribuição bastante acentuada da temperatura exterior. Este fenómeno reflete bem a grande variação da temperatura existente ao longo dos meses de inverno. É visível ainda que as linhas apresentam uma configuração muito semelhante indicando uma possível grande proximidade territorial das diferentes habitações.

Passando agora a análise para as humidades podemos observar na Fig. 4.6 a variação da média diária da temperatura durante os meses de inverno.

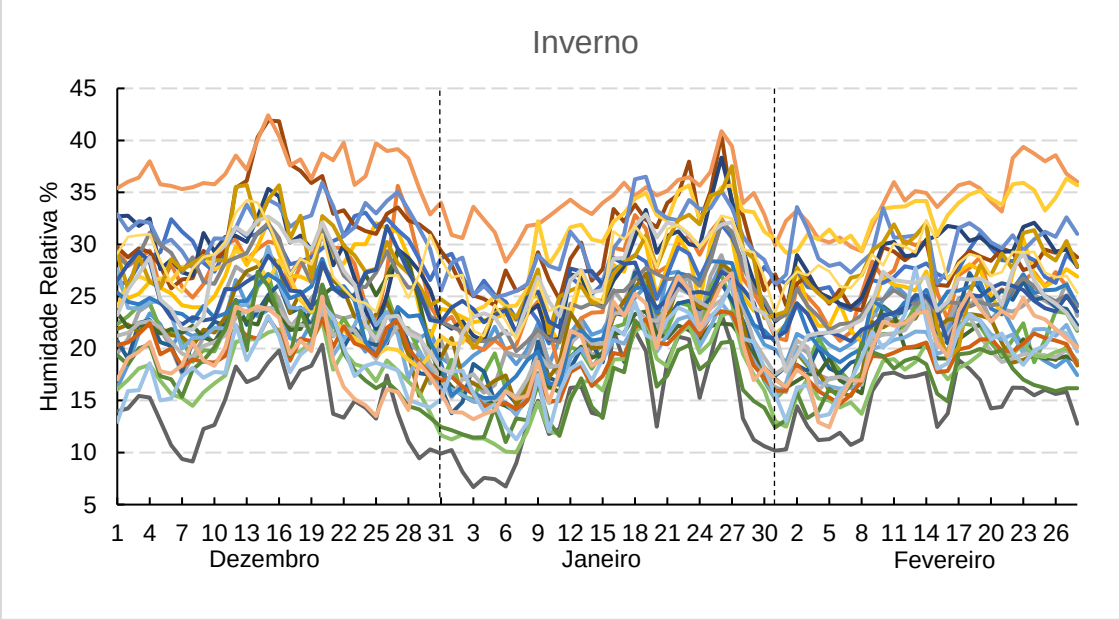


Fig. 4.6. Variação da humidade relativa média diária no inverno.

A variação ao longo do tempo observada segue, regra geral, a configuração apresentada nas anteriores figuras (Fig. 4.3 e Fig. 4.4). Essa semelhança verifica se inclusive na existência de "picos" mais baixos de humidade ao fim de semana com um aumento ao longo da semana. É observável que apesar dos picos existentes ao longo das curvas, estas apresentam tendencialmente valores mais baixos no início de cada mês, isto é, o intervalo de valores sobre uma translação mantendo aproximadamente os mesmos extremos. De todas as habitações aqui apresentadas, existem duas que apresentam uma configuração diferente de todas as outras. Essas curvas são novamente aquelas que foram já descritas como possíveis *outliers* desta amostra. É ainda de notar a habitação marcada a castanho.

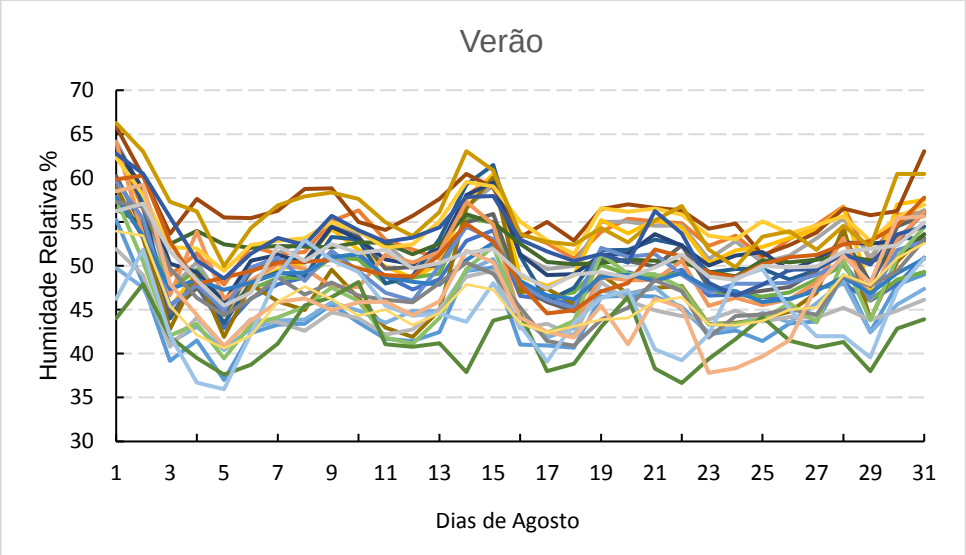


Fig. 4.7. Variação da humidade relativa média diária no verão.

Observando a Fig. 4.7, com a variação da humidade ao longo do verão denota existir uma variação cíclica, isto é, apresenta valores mais altos no início, a meio e no final do mês. Também se verifica, existirem humidades mais baixas ao fim de semana comparativamente ao resto da semana. Por fim é identificável apenas na curva a verde uma variação diferente da normalidade, representado esta a mesma habitação que se identifica como possível *outlier* nas figuras Fig. 4.3, Fig. 4.4 e Fig. 4.6.

Apresenta-se ainda a Fig. 4.8 com a variação da humidade relativa exterior ao longo dos meses de inverno. De uma observação mais geral é possível verificar a teoria da proximidade das habitações também está presente na distribuição da humidade visto que as habitações apresentam variações semelhantes. Observa-se a existência de humidades relativas altas a maior parte do tempo. É ainda visível que as habitações apresentam uma variação semelhante ao longo do inverno. Enquanto nas medições interiores é possível distinguir mais facilmente características individuais de algumas habitações, nas medições exteriores o mesmo não se verifica como é compreensível.

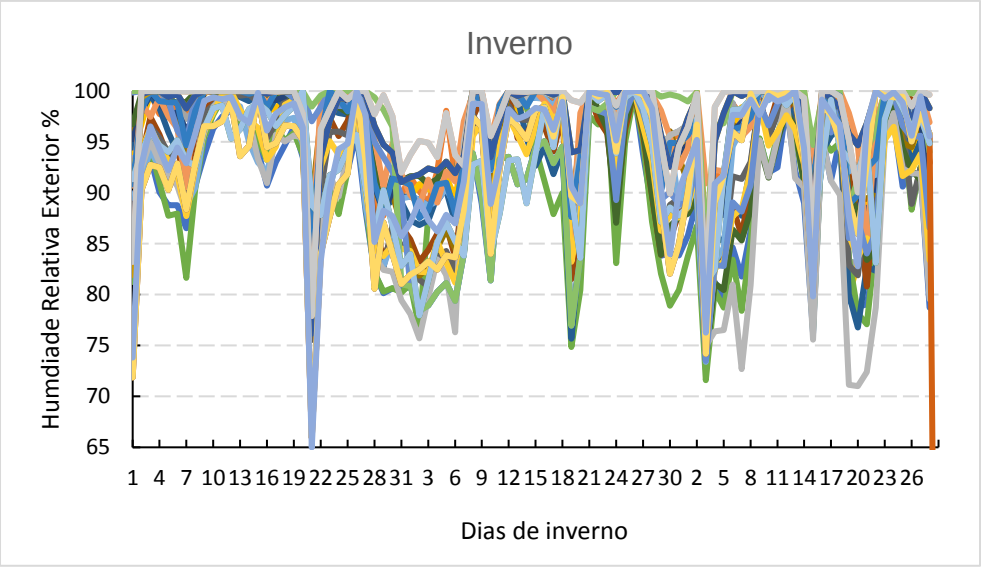


Fig. 4.8. Variação da humidade relativa média diária exterior nos meses de inverno.

Ao longo da análise anterior foi possível retirar diferentes conclusões para este estudo. A primeira é que os meses que representam o inverno contêm uma maior homogeneidade e consistência entre si ao contrário do que acontece com o verão, como já foi analisado. Estes fenómenos levaram a que o verão fosse caracterizado apenas pelo mês de agosto e o inverno pela totalidade dos seus três meses. O fato de dezembro, janeiro e fevereiro serem meses bastante coesos abre a possibilidade de os aglomerar numa única componente chamada inverno. A partir deste momento existiram apenas dois grupos de análise, o grupo de inverno e o grupo de verão.

O passo seguinte nesta fase foi a comparação a um nível geral tentando perceber a coerência dos resultados. Os dados são aqui visualizados através de medidas estatísticas cuja função é descrever a amostra.

Tabela 4.1. Descrição estatística da temperatura e humidade relativa no período de inverno e no verão.

Inverno			Verão		
	Humidade	Temperatura		Humidade	Temperatura
Média	23,8	21,4	Média	49,5	24,8
Desvio	4,8	1,3	Desvio	4,0	1,1
Máximo	36,8	24,4	Máximo	58,4	27,2
Mínimo	14,2	16,7	Mínimo	39,5	22,3
C. Assimetria	0,309	-1,097	C. Assimetria	-0,1	-0,097
C. Variação	0.202	0.063	C. Variação	0.081	0.045

Analisando a Tabela 4.1 podem-se retirar várias conclusões à cerca destes dois grupos de dados a analisar. Observando a temperatura média podemos dizer que esta não varia muito, contudo esta situação não se mantém relativamente às humidades, que apresentam um valor superior durante o verão. Saltando os extremos presentes na Tabela 4.1 é ainda possível observar que em termos de assimetria no inverno existe uma assimetria muito diferente comparando a humidade com a temperatura. A primeira é caracterizada por uma assimetria positiva (ligeiramente à direita) e a segunda uma assimetria negativa (fortemente à esquerda). Relativamente às humidades pode-se dizer que esta apresenta um coeficiente de assimetria baixo, que indica que a sua distribuição não se afasta muito da distribuição Normal. A assimetria presente nas temperaturas este é de grandeza negativa elevada, ou seja, a maior parte da distribuição está acima da média tendo uma maior concentração de temperaturas nas gamas mais elevadas. O período de verão apresenta uma distribuição praticamente normal por parte da temperatura e humidade relativa.

Pelos coeficientes de variação é possível verificar a grande homogeneidade existente de toda a amostra em ambos os períodos, com a exceção das humidades no inverno. Este fenómeno claramente corresponde a alguma anormalidade existente na informação.

Continuando o estudo da distribuição do conjunto de dados em análise apresentam-se em seguida uns gráficos do tipo "*Box-Plot*" sobre a humidade relativa e temperatura de modo a conseguir uma imagem mais clara sobre esta matéria.

Na Fig. 4.9 são apresentados os diferentes *Box-Plots* para a amostra de dados em estudo. Neles podemos ter uma visualização mais concreta de tudo que aqui foi dito. É ainda de notar que na figura do canto superior esquerdo apresentam-se 2 *outliers* correspondentes às habitações 2014 e 2008. Estas habitações correspondem às curvas enunciadas como possíveis *outliers* na descrição das figuras Fig. 4.3, Fig. 4.4 e Fig. 4.6. Por fim conclui-se que as distribuições são praticamente normais e que o coeficiente apresentado na Tabela 4.1 relativo à assimetria induz em erro relativamente à realidade, sendo este causado pela presença dos *outliers*. Este fato assegura ainda mais a equivalente importância que a visualização gráfica contém.

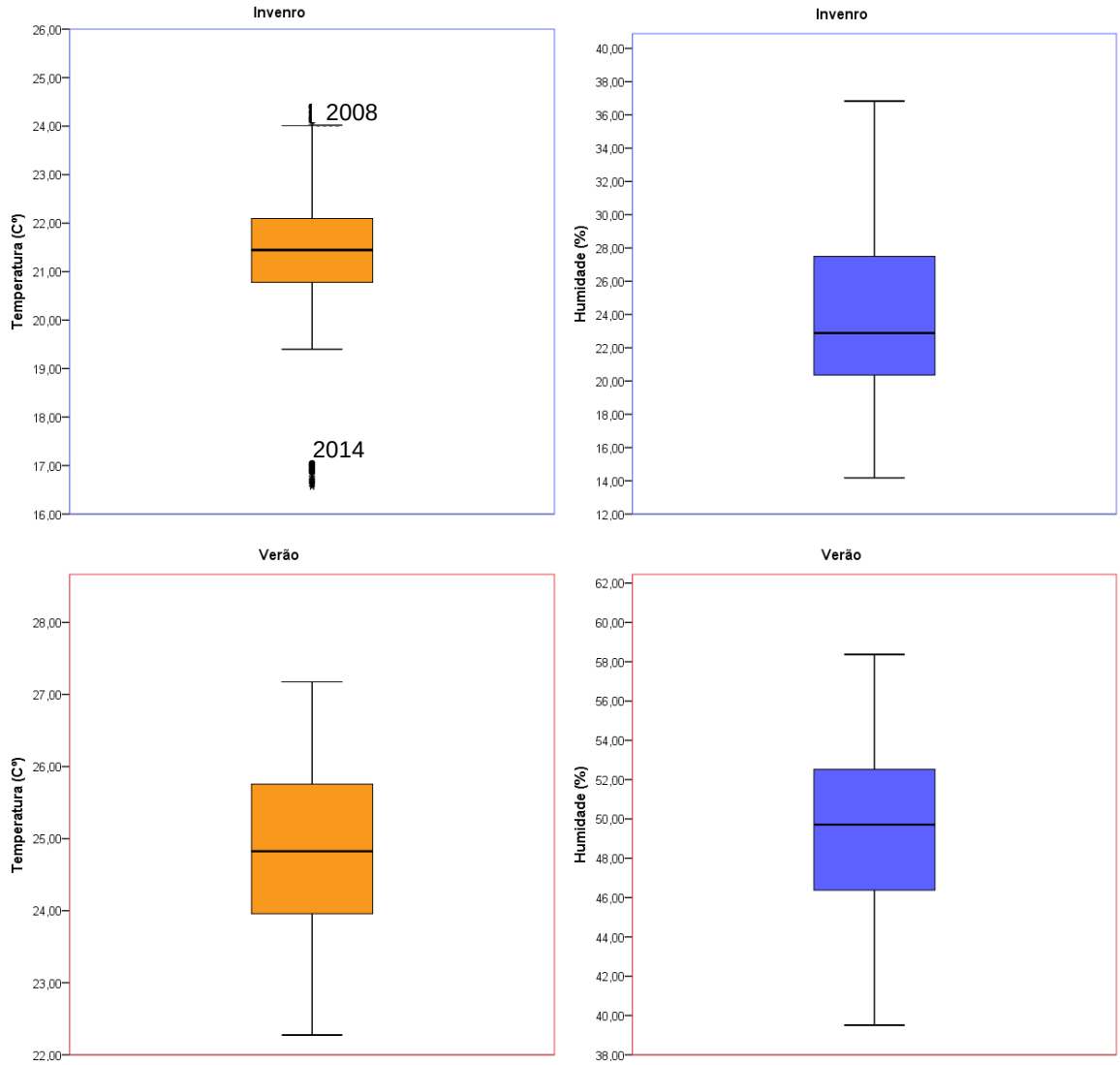


Fig. 4.9. Representação da distribuição dos dados em "*Box-Plots*".

Após esta descrição ficou a faltar perceber em que ponto se encontravam os *outliers* presentes na Fig. 4.9. Então de modo a tentar sintetizar toda a informação foi elaborada uma análise correspondente às médias diárias do período de inverno. Essa análise teve como resultado o gráfico de dispersão seguinte (Fig. 4.10).

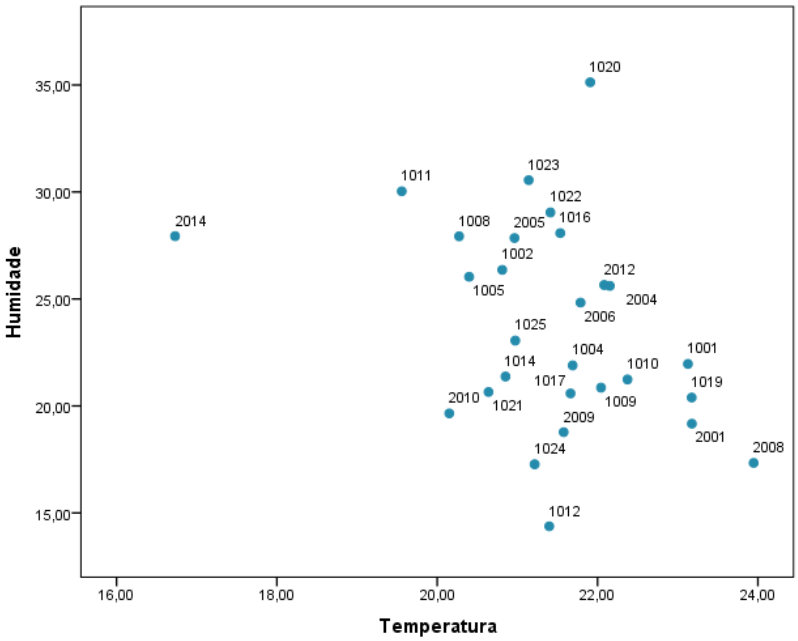


Fig. 4.10. Gráfico de dispersão com médias diárias.

Na Fig. 4.10 observa-se a dispersão das diferentes habitações de um modo generalizado, comparando a sua temperatura com a humidade relativa. A Fig. 4.9 indica a presença de duas habitações do tipo *outliers*, 2008 e 2014, contudo nesta dispersão generalizada apenas a habitação 2014 fica representada como claro *outlier*. Por esse motivo foi decidida a exclusão isolada da habitação 2014.

4.2. TRANSFORMAÇÃO E REDUÇÃO DE DADOS

Apresentam-se nesta secção os resultados obtidos após a transformação e redução do conjunto de dados. Em 4.1. foi apresentada uma descrição estatística detalhada sobre a base de dados de que dispomos. Como se observou, o conjunto de dados em estudo sofreu uma ligeira redução das suas constituintes. Essa redução não fica nesta fase apresentada pois surge como uma consequência da seleção e formatação da informação para que possa ser estudada. Ao contrário da redução em causa, a que aqui se encontra surge graças a um processo específico para o efeito e não como produto secundário de um outro processo.

Até ao momento a base de dados apresenta um número de variáveis considerável. Apesar da redução do número de casas a estudar; da consideração conjunta dos meses de inverno e apenas o mês de agosto como representativo do verão; e do cálculo das médias e desvios padrões horários; a quantidade de dados continua a ser excessiva para uma boa análise de *Data Mining*. Para que o leitor tenha uma melhor noção da quantidade de informação até ao momento, o número de variáveis corresponde a um total de $24 * 2 * 2 * 2 = 192$ variáveis. O produto anterior resulta da consideração de 2 estações diferentes (inverno e verão), com cálculo de médias e desvios padrões horários (24) para a temperatura e humidade relativa.

Após algumas considerações sobre número de variáveis atual, confirma-se a inevitabilidade da sua redução. Para tal procedeu-se à análise de correlação existente entre as diferentes variáveis. Esse estudo consistiu no cálculo de uma matriz de correlação para que fosse possível visualizar quais as variáveis que teriam uma forte relação entre si. Nas que fosse o caso essas poderiam ser transformadas reduzindo assim o conjunto de dados.

As matrizes de correlação revelaram que as variáveis encontravam-se praticamente todas correlacionadas. Este fenómeno era mais evidente nas variáveis do mesmo tipo, isto é, as médias das temperaturas eram mais correlacionadas com as médias da humidade relativa, tendo um comportamento semelhante para os desvios padrões. Na Fig. 4.11 é apresentada a matriz de correlação obtida para os meses de inverno. Nela observam-se 3 cores diferentes onde demonstram diferentes graus de correlação (vermelha mais forte, amarela mais fraca). Pode-se também observar que existem correlações direta e inversamente proporcionais. Esta problemática levou a uma forçosa redução do número de variáveis através da análise dos componentes principais. Esta fase é bastante importante para a aplicação da análise de *clusters*, porque a análise de *clusters* não deve ser aplicada a um conjunto de variáveis correlacionadas.

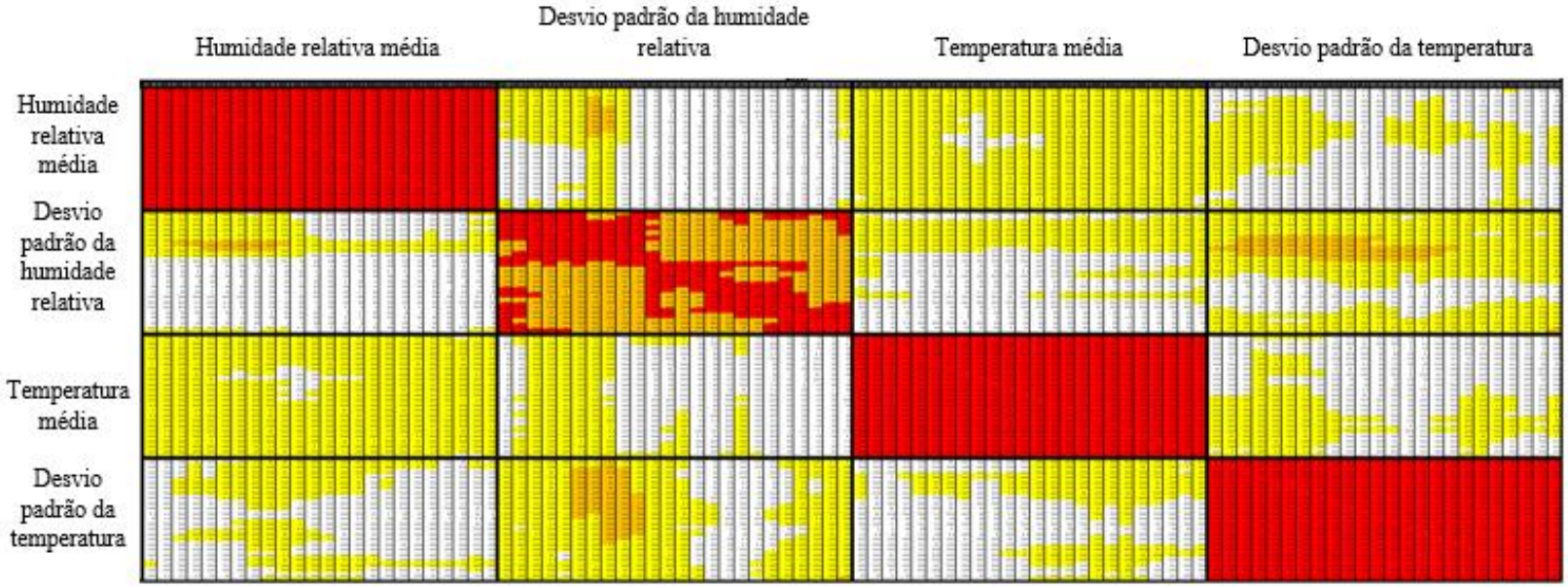


Fig. 4.11. Matriz de correlação respetiva aos meses de inverno.

A análise dos componentes principais foi responsável por reduzir consideravelmente o número de variáveis para um total bastante mais fácil de estudo. Os resultados obtidos foram consequência de um conjunto de tentativas para tentar descobrir qual o melhor conjunto de dados a considerar na ACP. A cada estação foi aplicada a ACP para a totalidade das 96 variáveis de inverno e 96 varáveis de verão. O resultado será um conjunto de componentes principais de inverno e outro de verão (conjuntos de variáveis independentes).

Tabela 4.2. Variância das principais componentes de inverno.

Componentes	Total	Variância (%)	Variância Acumulada (%)
1	35,759	37,249	37,249
2	21,376	22,266	59,515
3	16,690	17,385	76,900
4	13,465	14,026	90,926
5	3,552	3,700	94,626
6	1,853	1,930	96,556
7	0,739	0,770	97,326

Tabela 4.3. Variância das principais componentes de verão.

Componentes	Total	Variância (%)	Variância Acumulada (%)
1	46,472	48,409	48,409
2	27,255	28,391	76,799
3	10,602	11,044	87,843
4	3,283	3,420	91,263
5	2,359	2,457	93,720
6	1,882	1,961	95,681
7	1,067	1,111	96,792
8	0,689	0,717	97,509

A Tabela 4.2 e Tabela 4.3 apresentam o número ideal de componentes obtido no início da aplicação da ACP realizado através da análise de fatores do *software* SPSS. O número correspondente à última componente cujo total é superior à unidade, é considerado o número ideal de componentes. Pode-se observar que em ambas as situações a variância explicada por elas é de superior a 96%. Este valor garante a fiabilidade destas componentes na descrição da amostra inicial, pois encontra-se dentro das margens do aceitável (96% a 98%). Deste modo foram calculados os *scores* das 6 componentes principais do período de inverno e das 7 componentes principais do período de verão. A partir deste

momento o restante processo de DM será aplicado aos *scores* obtidos das respetivas componentes principais.

4.3. ANÁLISE DE CLUSTERS

Neste subcapítulo apresentam-se os resultados respetivos à análise de *Clusters*. É a partir deste ponto que a análise de DM começa a afirmar-se, sendo que até aqui tratava-se maioritariamente de descrição e tratamento de dados.

Nesta fase são utilizadas componentes obtidas em 4.2, agrupando ou isolando as diferentes habitações (*clusters* e *outliers*, respetivamente*)* segundo a sua similaridade, isto é a distância euclidiana entre os objetos e o centroide do *cluster*. Como visto em 2.2.2, um dos principais problemas que esta fase apresenta é a determinação do melhor número de *clusters*. Para tal aplicou-se um método hierárquico para a análise de *clusters* com o intuito de perceber qual seria o número adequado de *clusters* para uma posterior aplicação num método não-hierárquicos. O resultado obtido foi fruto de várias experiências, definindo diferentes limites de similaridade, até encontrar um resultado convergente.

Procedeu-se a aplicação dos métodos hierárquicos de *Complete Linkage* e *Ward's* para o estudo sobre o número ideal de *clusters*. Foram aplicados ambos os métodos com o intuito perceber, com base numa análise de várias perspetivas, a coerência dos resultados. Na Fig. 4.12 apresenta-se o dendrograma obtido por aplicação do método *Complete Linkage*.

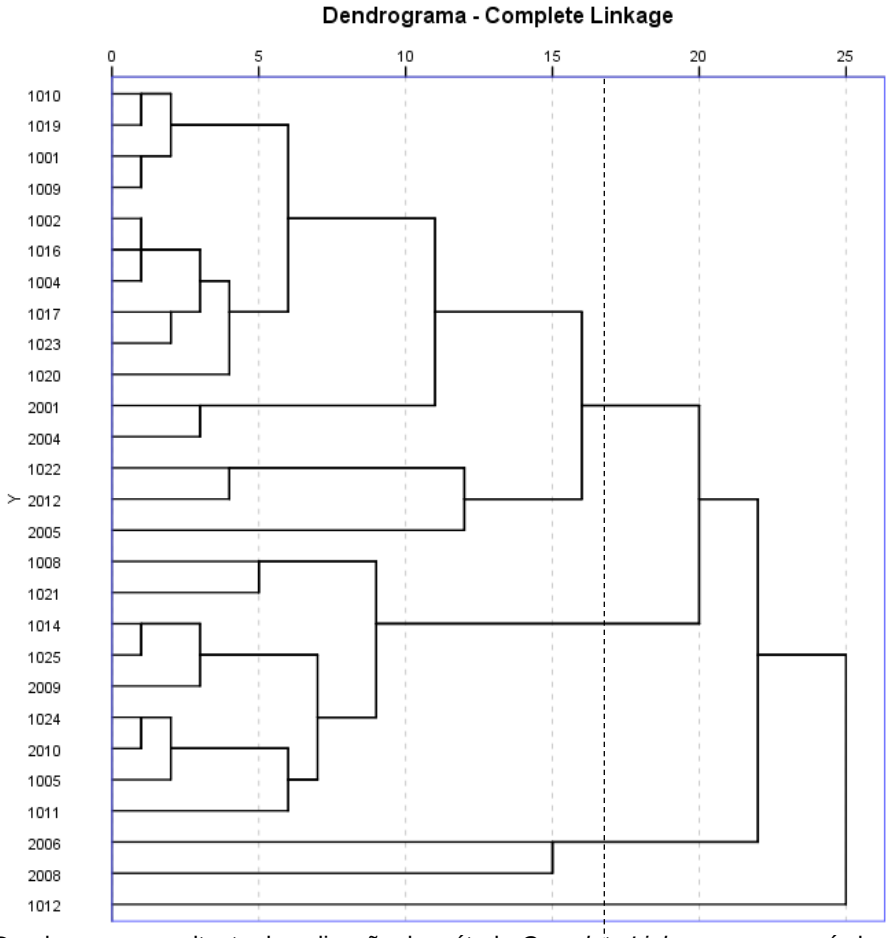


Fig. 4.12.Dendrograma resultante da aplicação do método *Complete Linkage* para o período de Inverno.

A linha a tracejado representa a divisão escolhida para a seleção do número de *clusters* neste período de inverno. Como é possível observar esta linha distingue dois *clusters*, de grande dimensão, e 3 pequenos, considerados *outliers*, fazendo um total de 5 grupos. Note-se que apesar de esta linha não individualizar os três *outliers* (consideram-se também *outilers* individuais as habitações 2006 e 2008), estes foram assim considerados devido a já referida análise de diferentes perspetivas, através de várias aplicações das ferramentas disponiveis. Com os vários cortes testados, experimentando com ambos os métodos e comparando os resultados no algoritmo *k-means* (em que k representa o número de cortes) foi possível concluir que as habitações 2006, 2008 e 1012 são *outliers*.

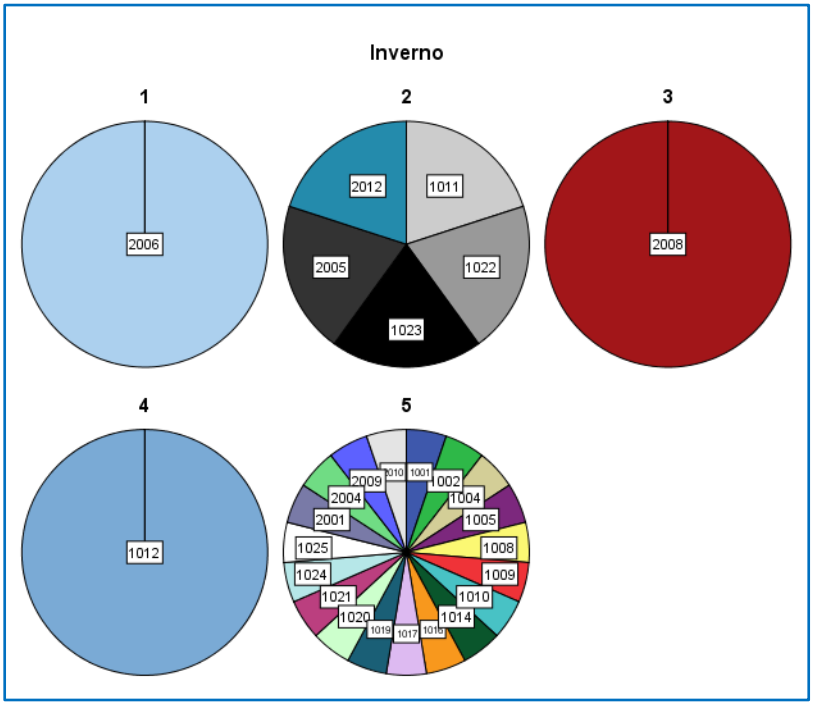


Fig. 4.13. Resultado do algoritmo de análise de *clusters K-means*, em que K=5.

A Fig. 4.13 apresenta o resultado do algoritmo *K-means* em concordância com o dendrograma anterior. São visíveis 5 grupos independentes de casas, graças à consideração de k=5, o número de *clusters* a formar mais coerente tendo em conta todo o processo iterativo realizado. Observam-se novamente 3 *outliers*, as habitações 2006, 2008 e 1012. É ainda de reparar que a agregação nos diferentes grupos não é totalmente coerente (apenas duas habitações não coincidem com a separação demonstrada no dendrograma). Este fenómeno é explicado por duas razões, em que a segunda é uma consequência da primeira. A primeira é que os algoritmos presentes por detrás de ambos os métodos (*K-means* e *Complete Linkage*) sofrem pequenas variações. A segunda razão é a baixa disparidade da amostra como já constado na secção 4.1. Sendo as temperaturas e humidades relativas muito pouco dispersas uma pequena alteração na forma do cálculo das distâncias (similaridade) pode resultar em resultados diferentes. Estas pequenas variações não apresentam uma grande preocupação visto que este tipo de análises não dispõem de um único resultado correto, pois dependem da sensibilidade do analista e da conformidade de resultados dos diferentes testes realizados.

Passando a análise de *clusters* ao mês de agosto foi necessária uma abordagem diferente. O processo de determinação dos diferentes *clusters* iniciou-se como apresentado até agora. Contudo após experimentar várias linhas divisórias semelhantes à da Fig. 4.12, o resultado não era muito conclusivo. Esta indeterminação deve-se à incapacidade dos diferentes *k-means* formarem grupos minimamente consistentes com as diferentes divisões testadas nos métodos hierárquicos. Contudo ao longo das diferentes tentativas começou a ser visível um padrão interessante. As habitações que aparentavam ser *outliers* eram praticamente as mesmas apenas distribuídas de formas diferentes. Então selecionaram-se as habitações indicadas presentes nos métodos hierárquicos que apresentavam característica forte de *outlier* e cruzou-se essa informação com os diferentes resultados de *k-means.*

Tabela 4.4. Seleção do número de *clusters*.

Habitações	2-*means*	3-*means*	4-*means*	5-*means*	6-*means*	Frequência
1001	✓	✓		✓		3
1014	✓		✓	✓	✓	4
1017	✓	✓	✓	✓	✓	5
1019		✓	✓	✓	✓	4
1021	✓	✓		✓		3
1024	✓	✓	✓		✓	4
2008				✓	✓	2
2009	✓				✓	2
2010	✓	✓	✓	✓		4

Após o conhecimento das frequências em que cada habitação aparece como *outlier* para os diferentes resultados de *K-means* eliminaram-se por ordem decrescente de frequência os algoritmos que não apresentavam as habitações mais frequentes. Com isto o algoritmo selecionado foi o 4-*means*,notando-se que os números de k escolhidos para comparação eram em todos os casos aglomerações possíveis e coerentes com os possíveis *outliers* identificados nos métodos hierárquicos.

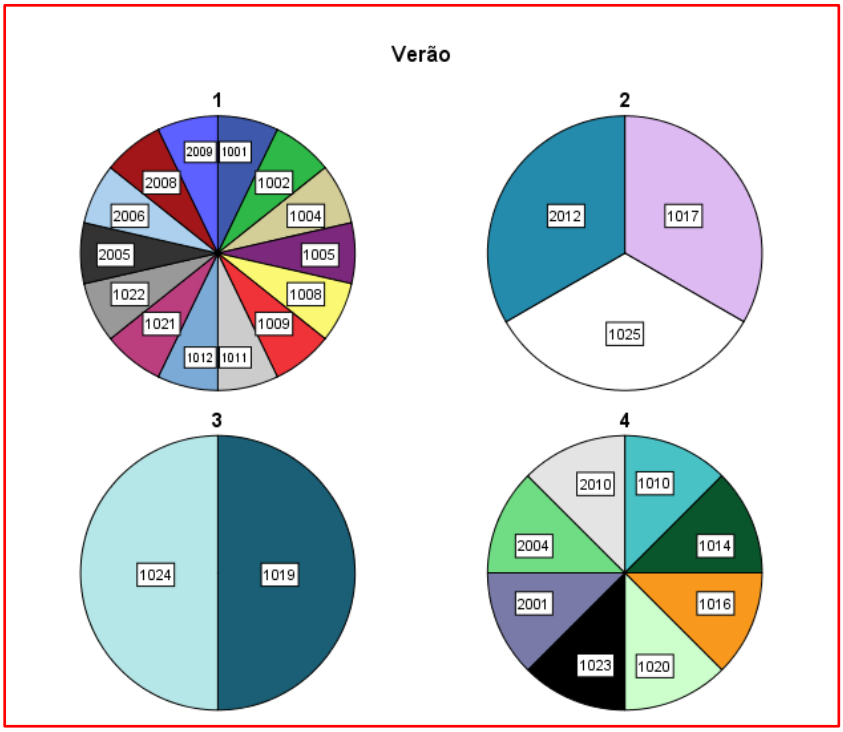


Fig. 4.14. Resultado do algoritmo de análise de *clusters* K-means, com K=4.

O resultado levado a cabo foi aquele que se apresenta na Fig. 4.14. Os dados relativos ao verão apresentaram este impasse devido ao facto da sua dispersão ser ainda menor que a presente durante o inverno. Esse facto pode ser confirmado pelos baixíssimos valores do coeficiente de variação apresentados na Tabela 4.1 e pelos *Box-Plots* ilustrados. A dificuldade na obtenção dos *clusters*, demonstra ainda forte ligação entre as diferentes variáveis, o que se verificará prejudicial em fases mais avançadas deste processo de DM, aquando a tentativa de identificar padrões na eficiência energética. Tendo em conta que os gastos energéticos são mais influenciáveis durante períodos de inverno, este fenómeno é compreensível, e por essa razão a partir deste ponto apenas se consideram os meses de inverno na padronização da amostra relativamente à eficiência energética.

A Fig. 4.15 apresenta uma síntese dos resultados obtidos através dos *clusters*. Observa-se um gráfico que compara as médias horárias das humidades relativas e temperaturas de cada grupo de habitações. As formas geométricas, assim como o número adjacente, identificam o *cluster* respetivo. A diferente coloração e tamanho de cada *cluster* representa as médias dos desvios padrões horários das humidades relativas e temperaturas de cada grupo de habitações, respetivamente.

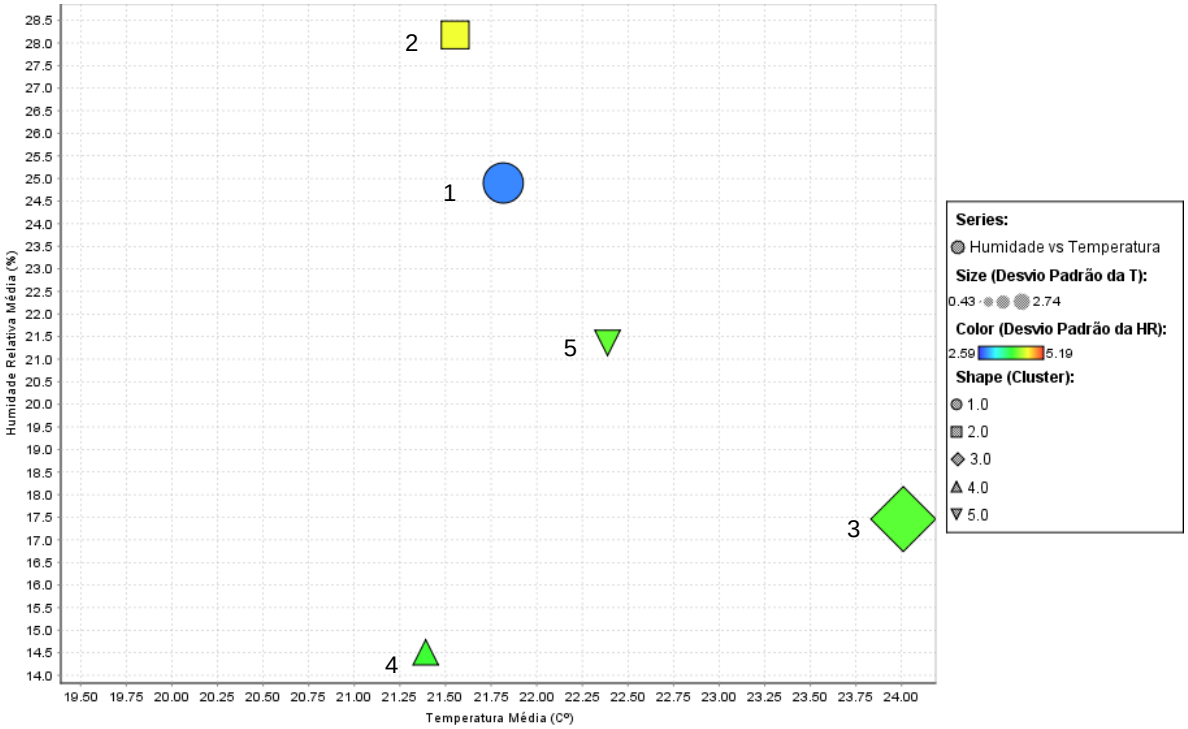


Fig. 4.15. Gráfico síntese dos *clusters* formados no período de inverno.

Através da Fig. 4.15 é possível atribuir uma caracterização rápida a cada *cluster*. Por outras palavras, podemos identificar que o *cluster* 2, apresenta maior humidade relativa e maior desvio padrão. O *cluster* 4 apresenta a humidade relativa mais baixa da amostra contudo o seu desvio padrão encontra-se dentro da média do conjunto. À exceção do *cluster* 3, o valor da temperatura média é aproximadamente igual para todos os *clusters*. Identifica-se que os *clusters* 4 e 5 apresentam o menor desvio padrão, seguindo-se os *clusters* 1 e 2 e por fim o *cluster 3*, com o maior desvio padrão.

4.4. PARAMETRIZAÇÃO DE EDIFÍCIOS

A parametrização de edifícios diz respeito à sua classificação segundo as variáveis qualitativas enuncidas em 3.2. Este passo é necessário para que possa ser construída a árvore de decisão, criando assim o elo de ligação entre o início da árvore e a classificação final, identificando assim os padrões característicos do processo de DM.

A Tabela 3.1 apresenta as diferentes variáveis utilizadas, as de carácter contínuo e as qualitativas. Com a obtenção dos *clusters* em 4.3 ficaram tratadas as variáveis contínuas, ficando apenas necessário refletir sobre as restantes. Tendo em conta que, à exceção de possíveis dados em falta, a cada habitação corresponde um certo atributo de uma dada variável a sua análise passa pelo estudo da frequência acumulada dessa variável. O cálculo das frequências acumuladas transmitem o número de vezes que uma certa variável assume um respetivo valor ou abaixo dele. No presente trabalho a divisão classificativa de cada variável qualitativa irá ser realizada com base nas frequências acumuladas. Assumindo uma classificação dividida em 3 categorias com intervalos idênticos para as 3 variáveis, consumo energético, espessura do isolamento e RPH.

Os gráficos seguintes apresentam as frequências acumuladas das diferentes variáveis.

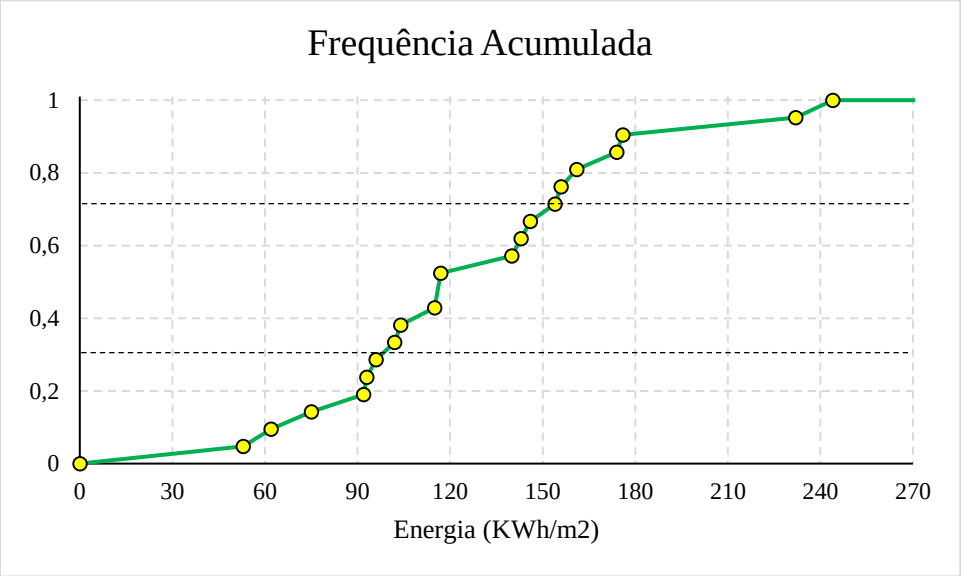


Fig. 4.16. Gráfico de frequências acumuladas da energia.

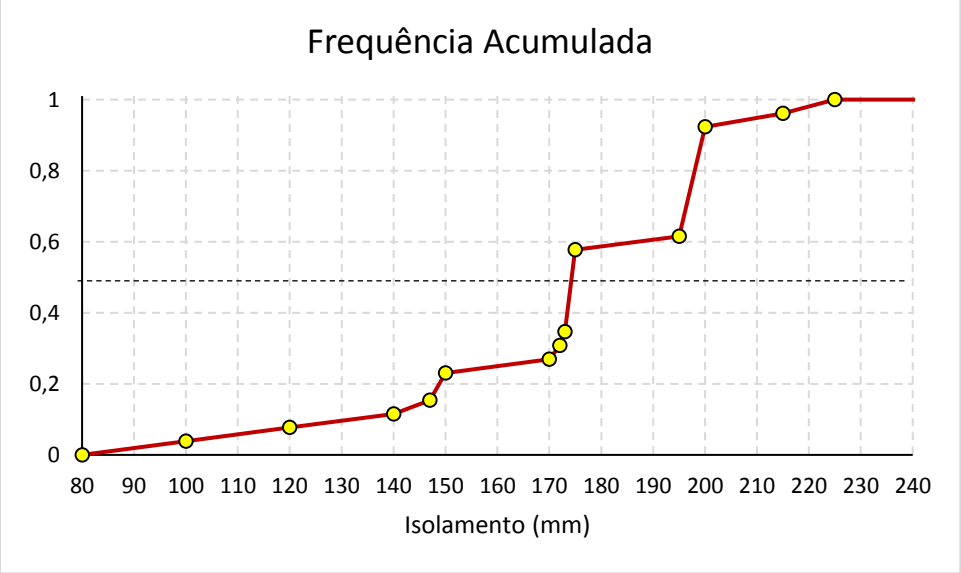


Fig. 4.17. Gráfico de frequências acumuladas do isolamento.

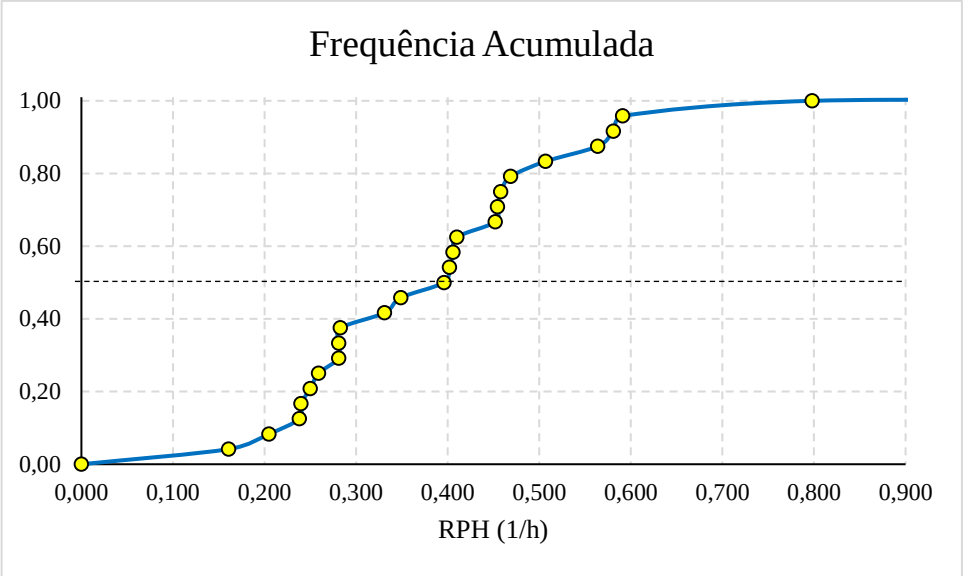


Fig. 4.18. Gráfico de frequências acumuladas do RPH.

Nas figuras anteriores observam-se também umas linhas a tracejado preto, representativas da divisão de classificações correspondentes tendo por base a frequência de ocorrência. Essas divisões seguiram o critério apresentado na Tabela 4.5.

Tabela 4.5. Resumo das classificações a atribuir.

Classificação	Energia	Classificação	Isolamento	Ventilação
>0.3	A	>0.5	I1	V1
0.3 a 0.7	B	<0.5	I2	V2
<0.7	C			

4.5. ÁRVORE DE DECISÃO

Definidas as diferentes classificações a atribuir é necessário agrupar as casas e reorganizá-las de modo a atingir o objetivo estabelecido. A Árvore de Decisão é responsável por agregar as diferentes classificações, com o intuito de classificar os diferentes grupos de habitações (*clusters)* em classes de energia anteriormente demonstradas.

Para a criação da árvore de decisão foi necessário a criação de um mapa de lógica. Este mapa serve de guia para a introdução ordenada dos dados, isto é, qual a ordem a que as varáveis irão estar sujeitas à medida que a classificação (desagregação) vai acontecendo. O primeiro nível de divisão corresponde aos resultados obtidos na análise de *clusters*. Já dentro do *cluster* respetivo identifica-se o primeiro classificador, o isolamento, que se divide em muito isolado (I1) ou pouco isolado (I2). O segundo nível de classificação corresponde ao RPH, com o número de renovações horárias alto (V1) ou baixo (V2). Por fim apresenta-se o classificador energia, para obtenção dos agrupamentos finais. Esta divide-se em 3 níveis de eficiência A (baixo consumo energético), B (consumo energético médio) e C (alto consumo

energético). Na Tabela 4.6 apresenta-se a identificação das diferentes habitações pertences aos diferentes *clusters* e respetivo nível de eficiência energética.

Tabela 4.6. Identificação das eficiências energéticas das habitações pertencentes a cada *cluster*.

Cluster	Eficiência Energética	Habitações
1	A	2006
2	A	1023
	B	2005-2012
	C	1022
3	C	2008
4	C	1012
5	A	1005-1020-1025-2010
	B	1009-1010-1016-1019-1024-2009-1001-1002
	C	1008-1014-1021-2001

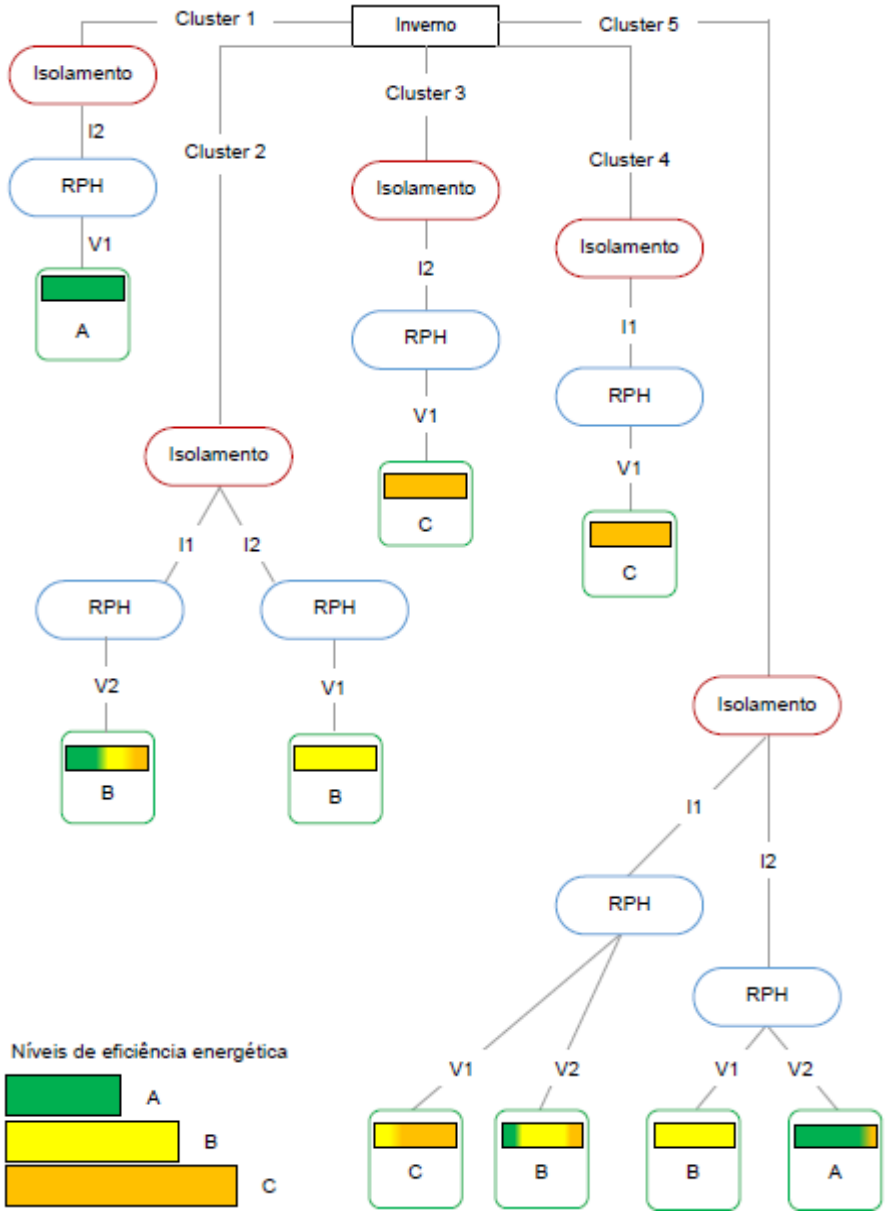


Fig. 4.19. Árvore de decisão.

Na Fig. 4.19 é possível observar a árvore de decisão construída. Ao percorrer a árvore observam-se os diferentes níveis de classificação até alcançar a eficiência energética. Esta é representada pela barra colorida estando a legenda apresentada na figura. A letra sob a barra respetiva representa o grau de eficiência energética que mais justiça faz na avaliação do conjunto habitacional dentro de cada derivação.

4.6. DISCUSSÃO

A determinação da árvore de decisão coloca fim à última fase de DM em que se aplicam métodos estatísticos e ou analíticos. Com o término dessa etapa fica apenas a faltar a interpretação dos resultados obtidos através de uma análise critica e cuidada.

Na avaliação da eficiência energética de habitações foram tomados em conta como variáveis de ponderação o isolamento, o RPH e o consumo energético de cada habitação de modo a conseguir classificar estas habitações em prol do conforto térmico interior. Existem várias abordagens possíveis relativamente à interpretação dos resultados obtidos e como tal optou-se por uma avaliação faseada aumentando os níveis de especificidade. Por outras palavras a discussão de resultados irá passar por considerar desde a generalidade dos *clusters* até aos tipos de isolamento e ventilação existentes nas habitações em estudo.

Começando por um nível mais geral de observação pode-se concluir que é possível determinar a eficiência energética do *cluster* com base na avaliação de cada derivação. Conclui-se que o *cluster* 1 representa o conjunto mais eficiente seguido do *cluster* 5, muito idêntico ao *cluster* 2. Por fim os *clusters* 3 e 4 apresentam a pior das classificações energéticas, o tipo C.

Uma vez avaliadas as classificações dos diferentes agrupamentos habitacionais é possível cruzar essa informação com a Fig. 4.15 e retirar algumas conclusões. Para humidades relativas mais baixas (*cluster* 3 e 4) verifica-se uma pior eficiência energética comparativamente com as habitações que contêm humidades mais altas (*cluster* 1,2 e 5). Até cerca de 25% da humidade relativa a eficiência energética aumenta, voltando a diminuir para valores superiores. Em termos de temperaturas verifica-se que para valores mais altos que a eficiência é menor (*cluster* 3).

Aumentando a especificidade da análise pode-se verificar que o isolamento tipo 1 é menos eficiente que o tipo 2, verificando-se o mesmo relativamente à ventilação, V1 pior que V2 na generalidade dos casos. Contudo as suas diferentes combinações indicam alguns padrões curiosos. É possível observar que a combinação I1+V1 é a combinação menos eficiente em termos energéticos. Já a combinação I2+V2 por sua vez apresenta boa eficiência energética. I1+V2 corresponde a uma combinação bastante moderada em termos energéticos (nível de eficiência de B). Por fim falta retratar a combinação I2+V1 que é das combinações mais fracas, contudo isto não se verifica no *Cluster* 1, o mais homogéneo em termos de humidade relativa. Esta incoerência pode estar ligada à distribuição que a relação entre a humidade relativa e a energia apresenta na Fig. 4.15, humidades relativas altas e baixas menor eficiência que as intermédias.

Com esta interpretação da árvore de decisão é possível identificar ainda algumas conclusões. Se elegermos uma análise isolada relativamente a cada componente é possível perceber que espessura do isolamento não é diretamente proporcional à eficiência energética, no exemplo, I1 apresenta nível de eficiência energética mais baixa que I2, verificando-se o mesmo para o RPH (V1 menos eficiente que V2). Sobre as diferentes combinações apenas se concluiu que I2+V2 é a mais eficiente e que I1+V1 é a menos eficiente. Uma possível causa para a contrariedade presente nestas duas análises (combinações ou componente a componente) é o diminuto número de habitações consideradas na árvore e a sua grande homogeneidade impedindo uma melhor distribuição.

5
CONCLUSÃO

5.1. CONCLUSÃO

Com a realização desta dissertação o conceito *Data Mining* pode ficar um pouco mais próximo da Engenharia Civil em Portugal. Colmatando as várias dificuldades e interpretando os resultados obtidos ao longo do processo de DM foi possível concluir que:

- *Data Mining* corresponde a um processo bastante intuitivo e apresenta uma grande variedade na escolha dos métodos a utilizar em cada fase;
- A análise de componentes principais é um método, de transformação e redução de dados, que se insere na metodologia *Data Mining*. Este método foi responsável por transformar as 192 variáveis inicias, 96 correspondentes aos meses de inverno e 96 ao verão, em 6 e 7 componentes principais, respetivamente, com uma caracterização das variáveis iniciais de aproximadamente 96%;
- A análise de *clusters* representa um método quase indispensável no processo de DM. Normalmente encontra-se associado à formação de grupos para futura classificação, mas também pode ser aplicado em fases precoces da metodologia. É ainda um dos passos mais exigentes de *Data Mining* visto que a determinação do número ideal de *clusters* é, sobretudo, um processo iterativo e com base na intuição do analista. Graças aos métodos de clusterização utilizados foi possível identificar 5 grupos de habitações quando se usou os dados relativos aos meses de inverno e 4 grupos para o verão;
- A árvore de decisão é um classificador fácil de elaborar, sem necessidade de grandes conhecimentos informáticos, e fácil de interpretar. Uma das desvantagens da utilização deste método é a quantidade de ramificações existentes. Se o número de níveis da árvore for elevado, a criação e interpretação desta poderá tornar-se difícil;
- O caso de estudo mostrou-se de difícil análise, devido essencialmente homogeneidade amostral. Apesar disso foi possível obter resultados e tecer algumas conclusões. As conclusões referem-se à amostra selecionada com informação completa;
- A grande quantidade de dados, normalmente requer a aplicação de *Data Mining* que exige um grau de conhecimento minimamente avançado sobre as ferramentas estatísticas a utilizar;
- Dos *softwares* escolhidos para o tratamento e análise dos dados, SPSS e Excel, a aplicabilidade dos métodos foi possível graças à grande compatibilidade entre eles. A capacidade de reprodução de resultados em SPSS só foi possível graças à abrangência do Excel, que colmatou algumas incapacidades do SPSS;
- Para a aplicação de processos de DM mais elaborados, que requerem a utilização de métodos analíticos mais avançados os programas de cálculo MATLAB e RStudio apresentaram maior liberdade.

Assim, os objetivos propostos foram atingidos e é possível concluir que, de um modo geral, *Data Mining* tem aplicabilidade no estudo da eficiência energética das habitações em relação ao conforto térmico. Essa aplicabilidade pode ainda ser potenciada através do uso de ferramentas mais complexas.

5.2. DESENVOLVIMENTOS FUTUROS

Graças à enorme adaptabilidade do método e liberdade que ele dispõe foi difícil controlar os diversos impulsos para experimentar e colocar à prova as diferentes opções disponíveis. A introdução de qualquer outra abordagem que não aquela explorada ao longo deste trabalho levaria a uma maior queda de rentabilidade e eficiência neste estudo. Com este fato tido em conta a melhor conclusão a retirar seria a sugestão de todas as opções como desenvolvimentos futuros para uma continuação de uma abordagem com grande potencialidade para a Engenharia Civil. São então sugeridas:

- A aplicação do processo aqui apresentado às diferentes bases de dados disponíveis,
- A comparação entre esta metodologia e a outra com objetivos idênticos;
- A utilização de outros *softwares* para o mesmo efeito e comparação de resultados;
- Aumento da especificidade da metodologia adotada, acrescentado algoritmo de classificação mais complexos.
- Estudar e demonstrar as diferentes aplicabilidades de *Data Mining* em várias etapas da construção, (projeto e execução).
- Aplicar *Data Mining* a diferentes processos de gestão e manutenção de edifícios.
- Execução de técnicas de DM associados a conceitos de cidades sustentáveis.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Han, Jiawei e Micheline Kamber. 2011. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc.

[2] García, Salvador, Julián Luengo e Francisco Herrera. 2015. *Data Preparation Basic Models*. Data Preprocessing in Data Mining. Cham: Springer International Publishing.

[3] Ren, Xiaoxin, Da Yan e Tianzhen Hong. 2015. "Data mining of space heating system performance in affordable housing". *Building and Environment* no. 89:1-13.

http://www.sciencedirect.com/science/article/pii/S0360132315000669.

[4] Usama, M. Fayyad, Piatetsky-Shapiro Gregory e Smyth Padhraic. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". Comunicação apresentada em THE SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, em Portland, Oregon.

[5] Fan, Cheng, Fu Xiao, Henrik Madsen e Dan Wang. 2015. "Temporal knowledge discovery in big BAS data for building energy management". *Energy and Buildings* no. 109:75-89.

http://www.sciencedirect.com/science/article/pii/S0378778815302991.

[6] Abdi, Hervé e Lynne J. Williams. 2010. "Principal component analysis". *Wiley Interdisciplinary Reviews: Computational Statistics* no. 2 (4):433-459. http://dx.doi.org/10.1002/wics.101.

[7] Sá, Joaquim P. Marques de. 2007. *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*. 2 ed.: Springer-Verlag Berlin Heidelberg.

[8] Witten, Ian H., Eibe Frank e Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques* Morgan Kaufmann Publishers Inc.

[9] Corporation, Oracle. 2014. "Oracle Database". http://www.oracle.com/index.html.

[10] Office, Microsoft. 1987. "Microsoft Office Excel". https://products.office.com/pt-pt/excel.

[11] Office, Microsoft. 1992. "Microsoft Office Access". https://products.office.com/pt-pt/access.

[12] IBM. 2008. "A history of progress". http://www.ibm.com/us-en/.

[13] Simon, Steve. 1998. "History of SPSS". *PMean*. http://blog.pmean.com/history-of-spss/.

[14] Software, DELL. 2016. "STATISTICA Product Index".

http://www.statsoft.com/Products/STATISTICA/Product-Index.

[15] MathWorks, The. "MATLAB".

http://www.mathworks.com/products/matlab/?requestedDomain=www.mathworks.com.

[16] RStuido. RStuido. https://www.rstudio.com/.

[17] RapidMiner. "RapidMiner Studio". https://rapidminer.com/.

[18] Waikato, University of. "Weka 3: Data Mining Software". http://www.cs.waikato.ac.nz/ml/weka/.

[19] Ahmed, A., N. E. Korres, J. Ploennigs, H. Elhadi e K. Menzel. 2011. "Mining building performance data for energy-efficient operation". *Advanced Engineering Informatics* no. 25 (2):341-354. <Go to ISI>://WOS:000290193700019.

[20] Fan, Cheng, Fu Xiao e Chengchu Yan. 2015. "A framework for knowledge discovery in massive building automation data and its application in building diagnostics". *Automation in Construction* no. 50:81-90. http://www.sciencedirect.com/science/article/pii/S0926580514002507.

[21] Chen, Qipeng, Zhong Fan, Dritan Kaleshi e Simon Armour. 2015. "Rule Induction-Based Knowledge Discovery for Energy Efficiency". *IEEE Access* no. 3:1423-1436.

http://dx.doi.org/10.1109/ACCESS.2015.2472355.

[22] Fan, Cheng, Fu Xiao e Shengwei Wang. 2014. "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques". *Applied Energy* no. 127:1-10. http://www.sciencedirect.com/science/article/pii/S0306261914003596.

[23] Motta Cabrera, David F. e Hamidreza Zareipour. 2013. "Data association mining for identifying lighting energy waste patterns in educational institutes". *Energy and Buildings* no. 62:210-216.

http://www.sciencedirect.com/science/article/pii/S0378778813001436.

[24] Yu, Z., F. Haghighat, B. C. M. Fung e H. Yoshino. 2010. "A decision tree method for building energy demand modeling". *Energy and Buildings* no. 42 (10):1637-1646. <Go to ISI>://WOS:000281417000008.

[25] Kim, Hyunjoo, Annette Stumpf e Wooyoung Kim. 2011. "Analysis of an energy efficient building design through data mining approach". *Automation in Construction* no. 20 (1):37-43.

http://www.sciencedirect.com/science/article/pii/S0926580510001044.

[26] Edwards, Richard E., Joshua New e Lynne E. Parker. 2012. "Predicting future hourly residential electrical consumption: A machine learning case study". *Energy and Buildings* no. 49:591-603.

http://www.sciencedirect.com/science/article/pii/S0378778812001582.

[27] Li, J. M., J. K. Ward, J. N. Tong, L. Collins e G. Platt. 2016. "Machine learning for solar irradiance forecasting of photovoltaic system". *Renewable Energy* no. 90:542-553. <Go to ISI>://WOS:000370102400050.

[28] Grolinger, Katarina, Alexandra L'Heureux, Miriam A. M. Capretz e Luke Seewald. 2016. "Energy Forecasting for Event Venues: Big Data and Prediction Accuracy". *Energy and Buildings* no. 112:222-233. http://www.sciencedirect.com/science/article/pii/S0378778815304461.

[29] Rodriguez Fernandez, M., A. C. Garcia, I. G. Alonso e E. Z. Casanova. 2016. "Using the Big Data generated by the Smart Home to improve energy efficiency management". *Energy Efficiency* no. 9 (1):249-260. <Go to ISI>://WOS:000368993200015.

[30] Ramos, Nuno M. M. e John Grunewald. 2015. *Annex 55 Reliability of Energy Efficient Building Retrofitting-Probability Assessment of Performance and Cost (RAP-RETRO) Stochastic Data*. Göteborg: Chalmers University of Technology.

[31] Kalamees, Targo, Juha Vinha e Jarek Kurnitski. 2005. "Indoor Humidity Loads and Moisture Production in Lightweight Timber-frame Detached Houses". *Building Physics No. 3* no. 29.

[32] Manelius, Elina e Juha Vinha. 2015. Air tightness, indoor climate and ventilation of Finnish timber framed single family houses. Annex 55 Reliability of Energy Efficient Building Retrofitting-Probability Assessment of Performance and Cost (RAP-RETRO) Stochastic Data.

[33] Kalamees, Targo, Minna Korpi, Juha Vinha e Jarek Kurnitski. 2009. "The effects of ventilation systems and building fabric on the stability of indoor temperature and humidity in Finnish detached houses". *Building and Environment* no. 44 (8):1643-1650.

http://www.sciencedirect.com/science/article/pii/S0360132308002540.

[34] Jokisalo, Juha, Jarek Kurnitski, Minna Korpi, Targo Kalamees e Juha Vinha. 2009. "Building-leakage, infiltration, and energy performance analyses for Finnish detached houses ". *Building and Environment* no. 44:377-387.